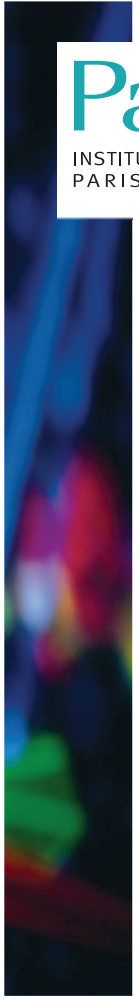
Statistical learning methods for ranking : theory, algorithms and applications

Sylvain Robbiano



ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2013-ENST-0033

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité "Signal et Image"

présentée et soutenue publiquement par

Sylvain ROBBIANO

le 19 juin 2013

Méthodes d'apprentissage statistique pour le ranking théorie, algorithmes et applications

Directeur de thèse: **Stéphan CLEMENCON**

Jury

M. Gabor LUGOSI, Professeur, Department of Economics, Pompeu Fabra University	Rapporteur
M. Massimiliano PONTIL, Professeur, University College London, UCL	Rapporteur
M. Francis BACH, Professeur, SIERRA, ENS Ulm	Examinateur
M. Arnak DALALYAN, Professeur, CREST, ENSAE Paristech	Examinateur
M. Nicolas VAYATIS, Professeur, CMLA, ENS Cachan	Examinateur
M. Stéphan CLEMENCON, Professeur, LTCI, TELECOM Paristech	Examinateur

TELECOM ParisTech
école de l'Institut Mines-Télécom - membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

THÈSE

RESUME:
Le ranking multipartite est un problème d'apprentissage statistique qui consiste à ordonner les observations qui appartiennent à un espace de grande dimension dans le même ordre que les labels, de sorte que les observations avec le label le plus élevé apparaissent en haut de la liste. Cette thèse vise à comprendre la nature probabiliste du problème de ranking multipartite afin d'obtenir des garanties théoriques pour les algorithmes de ranking. Dans ce cadre, la sortie d'un algorithme de ranking prend la forme d'une fonction de scoring, une fonction qui envoie l'espace des observations sur la droite réelle et l'ordre final est construit en utilisant l'ordre induit par la droite réelle.
Les contributions de ce manuscrit sont les suivantes : d'abord, nous nous concentrons sur la caractérisation des solutions optimales de ranking multipartite. Une nouvelle condition sur les rapports de vraisemblance est introduite et jugée nécessaire et suffisante pour rendre le problème de ranking multipartite bien posé. Ensuite, nous examinons les critères pour évaluer la fonction de scoring et on propose d'utiliser une généralisation de la courbe ROC nommée la surface ROC pour cela ainsi que le volume induit par cette surface. Pour être utilisée dans les applications, la contrepartie empirique de la surface ROC est étudiée et les résultats sur sa consistance sont établis.
Le deuxième thème de recherche est la conception d'algorithmes pour produire des fonctions de scoring. La première procédure est basée sur l'agrégation des fonctions de scoring apprises sur des sous-problèmes de ranking binaire. Dans le but d'agréger les ordres induits par les fonctions de scoring, nous utilisons une approche métrique basée sur le τ de Kendall pour trouver une fonction de scoring médiane. La deuxième procédure est une méthode récursive, inspirée par l'algorithme TreeRank qui peut être considéré comme une version pondérée de CART. Une simple modification est proposée pour obtenir une approximation de la surface ROC optimale en utilisant une fonction de scoring constante par morceaux. Ces procédures sont comparées aux algorithmes de l'état de l'art pour le ranking multipartite en utilisant des jeux de données réelles et simulées. Les performances mettent en évidence les cas où nos procédures sont bien adaptées, en particulier lorsque la dimension de l'espace des caractéristiques est beaucoup plus grand que le nombre d'étiquettes.
Enfin, nous revenons au problème de ranking binaire afin d'établir des vitesses minimax adaptatives de convergence. Ces vitesses sont montrées pour des classes de distributions contrôlées par la complexité de la distribution a posteriori et une condition de faible bruit. La procédure qui permet d'atteindre ces taux est basée sur des estimateurs de type plug-in de la distribution a posteriori et une méthode d'agrégation utilisant des poids exponentiels.

MOTS-CLES: Ranking Multipartite , Surface ROC , τ de Kendall, Arbres de Décision, Agrégation, Vitesses Minimax.

ABSTRACT:

Multipartite ranking is a statistical learning problem that consists in ordering observations that belong to a high dimensional feature space in the same order as the labels, so that the observations with the highest label appear at the top on the list. This work aims to understand the probabilistic nature of the multipartite ranking problem in order to obtain theoretical guaranties for ranking algorithms. In that framework, the output of a ranking algorithm takes the form of a scoring function, a function that maps the space of the observations to the real line which order is induced using the values on the real line.

The contributions of this manuscript are the following : first we focus on the characterization of the optimal solutions of multipartite ranking. A new condition on the likelihood ratios is introduced and shown to be necessary and sufficient to make the multipartite ranking well-posed. Then, we look at the criteria to assess the scoring function and propose to use a generalization of the ROC curve named the ROC surface. To be used in applications, the empirical counterpart of the ROC surface is studied and results on its consistency are stated.

The second topic of research is the design of algorithms to produce scoring functions. The first procedure is based on the aggregation of scoring functions learnt from bipartite sub-problems. To the aim of aggregating the orders induced by the scoring function, we use a metric approach based on the Kendall-τ to find a median scoring function. The second procedure is a tree-based recursive method inspired by the TreeRank algorithm that can be viewed as a weighted version of CART. A simple modification is proposed to obtain an approximation of the optimal ROC surface using a piecewise constant scoring function. These procedures are compared to the state of the art algorithms for multipartite ranking using simulated and real data sets. The performances highlight the cases where our procedures are well-adapted, specifically when the dimension of the features space is much larger than the number of labels.

Last but not least, we come back to the bipartite ranking problem in order to derive adaptive minimax rates of convergence. These rates are established for classes of distributions controlled by the complexity of the posterior distribution and a low noise condition. The procedure that achieves these rates is based on plug-in estimators of the posterior distribution and an aggregation using exponential weights.

KEY-WORDS: Multipartite Ranking, ROC Surface, Kendall-τ, Decision-Trees, Aggregation, Minimax Rates.

Contents

Introduction

On étudie dans cette thèse le problème de ranking à classes ordinales. Le ranking est une tâche qui consiste à ranger des observations dans l'ordre croissant de leur étiquette (inconnue) qui leur sont assignées, grâce à un ensemble d'exemples labellisés.
Supposons que l'on possède un ensemble d'observations, caractérisées par un ensemble de variables et d'une étiquette (ou d'une classe), appartenant à un ensemble discret ordonné. Un objectif habituel est de proposer une fonction qui prédit l'étiquette d'une nouvelle observation à partir des variables de l'observation. Cette tâche d'apprentissage supervisé est appelée régression ordinale. Cependant, il existe de nombreuses applications où le but est plutôt de définir un ordre sur l'ensemble des observations qu'une classification. Ce problème est appelé le ranking multipartite (ou ranking K-partite).
Dans de nombreuses situations, un ordre naturel peut être considéré sur un ensemble d'observations. En recherche d'information, le but est de ranger tous les documents par degré de pertinence pour une requête précise, à partir d'un ensemble d'entrainement décrivant les caractéristiques X d'un échantillon de documents et leur niveau de pertinence via une variable Y ordinale discrète, qui peut prendre plus que deux valeurs: dans le répertoire de données LETOR, elle prend cinq valeurs, allant de 0 ("non pertinent") à 4 "parfaitement pertinent". En médecine, les outils de prise de décision sont aussi requis dans le cadre multi-classes, les étiquettes correspondant à une gradation ordonnée de la maladie (de "sain" à "sérieusement malade") et les statistiques de test de diagnostic sont utilisées pour la discrimination entre les états pathologiques (cf. [Pepe, 2003], [Mossman, 1999] ou [Nakas & Yiannoutsos, 2004] par exemple).
Bien que ce problème soit omniprésent dans de nombreuses applications, des questions théoriques liées à l'apprentissage de fonctions de prédiction sont encore largement ouvertes. Plusieurs procédures ont été élaborées pour apprendre les fonctions de prédiction pour le ranking multipartite, mais la consistance des algorithmes n'a pas été abordée. La principale motivation de ce manuscrit est la compréhension de la nature probabiliste du problème ranking multipartite afin d'en déduire des algorithmes consistants. Ainsi, nous proposons deux méthodes pour atteindre cet objectif, la première en s'appuyant sur l'agrégation des fonctions de prédiction et le second basé sur un schéma d'approximation récursive.
En présence de réponse ordinale (*ie* le label Y en prenant un nombre fini de valeurs de $1, \ldots, K$, avec $K \geq 3$), la tâche ranking multipartite consiste à apprendre comment ordonner des observations non étiquetées de façon à reproduire le plus fidèlement possible l'ordre induit par les étiquettes qui ne sont pas encore observées. La façon naturelle de considérer ce problème consiste à construire une fonction de scoring s sur l'ensemble d'apprentissage qui donne une valeur réelle pour chaque observation

et utilise l'ordre naturel de la droite réelle pour ordonner les observations. Idéalement, quand la fonction scoring s augmente, avec une grande probabilité, nous nous attendons à observer majoritairement les observations avec l'étiquette $Y = 1$ en premier, celles avec l'étiquette $Y = 2$ ensuite, ... et les observations ayant le label $Y = K$ obtiennent les valeurs les plus élevées.
Le problème d'ordonner des données avec des étiquettes binaires, généralement appelés le *problème ranking binaire*, a récemment fait l'objet d'une grande attention dans la littérature statistique et de la machine-learning. Elle conduit à la conception de nouveaux algorithmes efficaces et bien adaptées à la tâche de ranking binaire (voir [Clémençon & Vayatis, 2009b], [Freund *et al.*, 2003] et [Clémençon & Vayatis, 2010] entre autres) et donne lieu à des développements théoriques importants dédiés à cet problème d'apprentissage global (voir [Agarwal *et al.*, 2005] ou [Clémençon *et al.*, 2008] par exemple). L'extension des concepts et des résultats liées au contexte du ranking multipartite est loin d'être immédiate et pose plein de questions de nature théoriques et pratiques, voir [Flach, 2004] et les références associées. Alors que, dans le cadre binaire, la courbe ROC (ainsi que les transformations et résumés de cette dernière telle que la célèbre aire sous la courbe ROC (AUC)) a fourni l'outil définitif pour évaluer la performance des règles de ranking depuis qu'elle a été introduite dans les années 40 (*cf* [Green & Swets, 1966]), ce n'est que récemment que cette mesure fonctionnelle de performance a été généralisée au cadre multipartite, conduisant à la notion de graphique ROC ([Scurfield, 1996]). Jusqu'à maintenant, l'approche suivie par la plupart des auteurs consistait à optimiser un critère scalaire précis sur un ensemble (non-paramétrique) de règles de ranking/scoring et appliquer la méthode de minimisation du risque empirique. Généralement, le risque de ranking compte le nombre de paires concordantes, c'est à dire le nombre d'observations qui sont rangées dans le même ordre que leurs étiquettes, et prend la forme d'une U-statistique de degré deux, voir [Clémençon *et al.*, 2008], [Rudin *et al.*, 2005]. Alternativement, dans le cadre binaire, cela peut être une fonction des rangs induits par la règle de ranking candidate, comme dans [Rudin, 2006], [Clémençon & Vayatis, 2007] ou [Clémençon & Vayatis, 2008].

Pour le ranking multipartite, diverses méthodes ont été proposées afin de développer des algorithmes efficaces. Nous pouvons regrouper ces méthodes en deux approches. Dans un premier groupe, on trouve des méthodes statistiques classiques, qui consistent à estimer les distributions a posteriori $\eta_k(x) = \mathbb{P}\{Y = k|X = x\}$ et utiliser ces estimateurs pour ordonner l'espace des observations. C'est le cas de l'analyse discriminante linéaire (LDA, [Fisher, 1936]) et la régression logistique (voir [Hastie & Tibshirani, 1990], [Friedman *et al.*, 1998], [Hastie *et al.*, 2001]) qui sont basées sur une maximisation de la probabilité dans un modèle de type logit. Un autre exemple est la régression logistique du noyau (voir par exemple [Zhu & Hastie, 2001]) qui permet d'estimer la distribution a posteriori par la résolution d'un problème d'optimisation convexe.

L'autre approche est basée sur l'optimisation d'un critère évaluant la performance empirique (ou le risque) d'une fonction de scoring. Parmi ces méthodes, on peut citer RankBoost proposée par [Freund *et al.*, 2003] et AdaRank proposée par [Xu & Li, 2007] qui sont fondées sur le principe du boosting (voir [Freund & Schapire, 1999]). Les méthodes RankSVM proposée par [Joachims, 2002] et RankRLS proposée par [Pahikkala *et al.*, 2007] qui sont basées sur SVM mais pour différentes fonctions de coût spécifiquement, la perte hinge et la perte quadratique. Toutes ces méthodes sont basées sur la même idée: l'optimisation d'un critère basé sur le nombre de paires concordantes. Ainsi, en fonction de la méthode, le critère de performance peut être écrit comme la somme pondérée d'AUC. Plus récemment, [Waegeman *et al.*, 2008a] a proposé un algorithme de type SVM pour optimiser un critère basé sur une perte de K-uplet d'observations qui peuvent être écrits en fonction des K-uplets discordants soit le nombre de K-uplets d'observations tels que l'ordre induit par la fonction de scoring n'est pas cohérent avec les étiquettes observées.
De toute évidence, ces deux approches présentent des avantages et des inconvénients. Bien qu'il semble naturel d'estimer les distributions a posteriori qui décrivent le modèle, cette approche a certaines limites. Tout d'abord, les représentations des distributions postérieures η_k utilisent des modèles de type logit qui peuvent ne pas être adaptés aux données observées. De plus, ces méthodes sont touchées par le fléau de la dimension lorsque l'espace des observations a une grande dimension. Dans ce dernier cas, les méthodes basées sur l'optimisation d'un critère empirique permettent d'obtenir des fonctions de scoring plus précises. Cependant, ces méthodes fournissent des fonctions qui sont bonnes pour le problème global mais ne peuvent assurer de trouver les meilleures observations d'un échantillon. Néanmoins, dans de nombreuses applications, telles que la recherche d'information par exemple, nous nous soucions seulement du haut de la liste. Ainsi, d'autres critères ont été introduits, tels que le gain cumulé actualisé (DCG voir [Cossock & Zhang, 2008], la précision moyenne (AP voir [Voorhees & Harman, 2005]) et le rang de réciprocité prévue (ERR) ([Chapelle & Chang, 2011]).

Dans ce manuscrit, nous caractérisons les solutions optimales du problème de classement et nous présentons un outil fonctionnel (la surface ROC) qui permet de retrouver les fonctions de scoring optimales. Contrairement à la tâche de ranking binaire, la tâche de ranking multipartite n'est pas un problème bien posé sans hypothèse supplémentaire. Nous utilisons la théorie classique de la monotonie stochastique ([Lehmann & Romano, 2005]) pour la rendre bien posée. Nous proposons plusieurs algorithmes basés sur la maximisation de la surface ROC qui ont les atouts des deux approches précédentes. Ces algorithmes produisent des fonctions de scoring qui imitent l'ordre induit par la fonction de régression η sans estimer directement la fonction de régression en grande dimension.

Les méthodes présentées dans ce manuscrit s'appuient sur la généralisation de la procédure de ranking binaire (voir [Clémençon & Vayatis, 2010],

[Clémençon *et al.*, 2013a]) au cadre du ranking multipartite. Toutes nos méthodes produisent des fonctions de scoring constantes par morceaux et peuvent être représentées par des arbres de décisions binaires orientés. Les feuilles représentent les cellules de la partition de l'espace des observations $\mathcal{X}$.

Liste de publications

Ces travaux ont été l'objet de plusieurs publications scientifiques :

- S. Clémençon and S. Robbiano, *Minimax Learning Rates for Bipartite Ranking and Plug-in Rules*, ICML 2011.
- S. Clémençon, S. Robbiano and N. Vayatis, *Ranking Data with Ordinal Label: Optimality and Pairwise Aggregation*, Machine Learning, 2013.
- S. Clémençon, S. Robbiano and J. Tressou *Maximal Deviations of Incomplete U-statistics with Applications to Empirical Risk Sampling*, SIAM data Mining, 2013.
- S. Robbiano *Upper Bounds and Aggregation in Bipartite Ranking*, EJS, 2013.

D'autres publications sont en cours de préparation :

- S. Robbiano, *Consistent Aggregation of Bipartite Rules for Ranking Ordinal Data*, soumis.
- S. Clémençon and S. Robbiano, *Confidence Regions for the ROC Surface via Smoothed Bootstrap*, soumis.
- S. Clémençon and S. Robbiano, *TreeRank Tournament*, soumis.

Résumé du chapitre 1

Le chapitre 1 est consacré à l'étude d'existence de fonction de scoring optimale pour le ranking à label ordinal et aux critères de performance. Dans le cas du ranking binaire cette question ne se pose pas, car il existe toujours une fonction de scoring optimale qui est la fonction de régression. Etant donnée cette propriété, on définit comme fonction optimale pour le ranking multipartite, toute fonction qui appartient à l'intersection des ensembles de fonctions optimales pour les sous-problèmes binaires. On montre que cette intersection est non vide, si et seulement si, une certaine condition sur les lois conditionnelles des classes est respectée. Cette condition est appelée monotonie des rapports de vraisemblance et est fortement relié à une autre condition existante dans la littérature (voir [Waegeman & Baets, 2011]). De plus, si cette condition est respectée, on a montré que les fonctions optimales sont les fonctions de scoring qui ordonnent dans le même ordre que la fonction de régression. On introduit ensuite la surface ROC qui est l'extension naturelle de la courbe ROC au cas à label ordinal. On a montré que les fonctions qui dominent toutes les autres en termes de surface ROC, sont exactement celles qui appartiennent à l'ensemble des fonctions optimales pour le ranking (sous l'hypothèse de monotonie des rapports de vraisemblance). On a ainsi un critère fonctionnel pour retrouver les fonctions optimales pour le problème du ranking. Ce critère n'étant pas aisé à manipuler, on a introduit le volume sous la surface ROC qui est donc un critère réel et qu'on utilise pour comparer de façon numérique les performances des algorithmes de ranking.

Optimalité et critères dans le contexte du ranking avec label ordinal

Cadre probabiliste et notations

On considère un système boite-noire avec une paire entrée/sortie aléatoire (X, Y). On suppose que le vecteur d'entrée X prend ses valeurs dans $\mathbb{R}^d$ et la sortie Y dans un ensemble discret et ordonné $\mathcal{Y} = \{1, \ldots, K\}$. On suppose que les valeurs de la sortie Y reflète un ordre sur $\mathbb{R}^d$. Le cas $K = 2$ est appelé ranking binaire. Dans les chapitre 1 à 6, on se concentre sur les cas où $K > 2$. On note ϕ_k la fonction de densité de la loi conditionnelle de X sachant $Y = k$ et par $\mathcal{X}_k \subseteq \mathbb{R}^d$ le support de ϕ_k. On pose également $p_k = \mathbb{P}\{Y = k\}$, $k = 1, \ldots, K$, le paramètre de mélange pour $Y = k$ et $\eta_k(x) = \mathbb{P}(Y = k \mid X = x)$ la loi a posteriori.

La fonction de régression $\eta(x) = \mathbb{E}(Y \mid X = x)$ peut être exprimée sous la forme suivante : $\forall x \in \bigcup_{l=1}^{K} \mathcal{X}_l$, $\eta(x) = \sum_{k=1}^{K} k \cdot \eta_k(x)$, comme l'espérance d'une variable aléatoire discrète. Pour la suite, on utilise par convention $u/0 = \infty$ pour tout $u \in]0, \infty[$ et $0/0 = 0$. $\mathbb{I}\{E\}$ est la fonction indicatrice de l'événement E.

Règles de scoring optimales

Le problème considéré dans cette thèse est d'inférer une relation d'ordre sur $\mathbb{R}^d$ après avoir observé un ensemble de données avec des labels ordinaux. Pour cela, on considère les règles de décision à valeurs réelles de la forme $s \ : \ \mathbb{R}^d \to \mathbb{R}$ appelée *fonctions de scoring*. Dans le cas des labels ordinaux, l'idée principale est que les bonnes règles de scoring s sont celles qui attribuent un score élevé $s(X)$ aux observations avec des grandes valeurs de labels Y. Une fonctions de scoring est dite optimale pour le problème du ranking à K classes si pour tout $k, l \in \{1, \ldots, K\}$, $l < k$, $\forall x, x' \in \mathcal{X}_l$, $\Phi_{k,l}(x) < \Phi_{k,l}(x') \Rightarrow s^*(x) < s^*(x')$.

L'idée dernière cette définition peut être comprise en considérant le cas $K = 2$. La classe $Y = 2$ doit obtenir des scores plus grand que la classe $Y = 1$. Dans ce cas, une fonction de scoring optimale s^* doit ranger les observations x dans le même ordre que la probabilité a posteriori η_2 de la classe $Y = 2$ (ou de façon équivalente au ratio $\eta_2/(1-\eta_2)$). Puisque $\eta_1(x)+\eta_2(x) = 1$, pour tout x, il est facile de voir que cette condition est équivalente à celle décrite dans [Clémençon & Vayatis, 2009b]. Dans le cas général ($K > 2$), l'optimalité d'une fonction de scoring s^* signifie que s^* est optimale pour tous les sous-problèmes binaires contenant les classes $Y = k$ et $Y = l$, avec $l < k$. Il est important de remarquer que, dans le cadre probabiliste introduit ci-dessus, une fonction de scoring optimale peut ne pas exister.

Existence et caractérisation des fonctions de scoring optimales

Notre premier résultat important est la caractérisation des lois pour lesquelles la famille des fonctions de scoring optimale n'est pas vide. Pour cela on introduit une hypothèse MLR (monotonie des rapports de vraisemblance) de monotonie des rapports de vraisemblance i.e. si pour un couple (k, l) tel que $k < l$ on a $\Phi_{k,l}(x) < \Phi_{k,l}(x')$ alors pour tout couple (k', l') tel que $k' < l'$ on a $\Phi_{k',l'}(x) < \Phi_{k',l'}(x')$. L'hypothèse MLR caractérise les lois de la paire aléatoire (X, Y) pour lesquelles le concept même de fonction de scoring optimale a du sens. Si cette condition n'est pas vérifiée alors la nature ordinale des étiquettes enfreinte. On signale qu'une condition, appelée *ERA ranking representability*, a été introduite dans [Waegeman & Baets, 2011], voir définition 2.1 . On donne ensuite plusieurs cas dans lesquels cette hypothèse est vérifiée, en particulier celui des familles exponentielles.

Mesures performance pour le ranking multipartite

Rappel sur le cas binaire

La courbe ROC (Receiver Operating Characteristic) est l'outil visuel de référence pour représenter les performances des tests statistiques cherchant à discriminer entre deux populations, voir [Green & Swets, 1966]. Ce graphique est très largement utilisé dans de nombreuses applications, comprenant le traitement du signal,

la recherche d'informations et l'examen de risque de crédit, voir [Fawcett, 2006]. En médecine par exemple, elle est utilisée pour évaluer les tests de diagnostiques, qui vise à discriminer les patients souffrant d'une maladie des autres à travers des mesures physicochimique ou l'occurrence possible de certains symptômes, voir [Pepe, 2003]. Dans notre cas, on est en présence de $K > 2$ labels, on va donc généraliser la courbe ROC pour l'adapter à notre cadre.

Surface ROC

Dans le chapitre 1, la surface ROC est définie pour K quelconque mais on se limite ici au cas $K = 3$. Soit une fonction $s : \mathbb{R}^d \to \mathbb{R}$, la surface ROC est un outil visuel qui reflète la façon dont les lois conditionnelles de $s(X)$ sachant que la classe $Y = k$ sont éloignées les unes des autres pour $k = 1, 2, 3$. On introduit la notation $F_{s,k}$ pour la fonction de répartition de la v.a. $s(X)$ sachant $Y = k$, $\forall t \in \mathbb{R}$, $F_{s,k}(t) = \mathbb{P}\{s(x) \leq t \mid Y = k\}$.

La surface ROC d'une fonction de scoring réelle s est définie comme le graphe de l'extension continue de la surface paramétrique sur le cube unité $[0, 1]^3$:

$$\begin{aligned} \Delta &\to \mathbb{R}^3 \\ (t_1, t_2) &\mapsto (\, F_{s,1}(t_1)\,,\; F_{s,2}(t_2) - F_{s,2}(t_1)\,,\; 1 - F_{s,3}(t_2))\ , \end{aligned}$$

où $\Delta = \{(t_1, t_2) \in \mathbb{R}^2 \ : \ t_1 < t_2\}$. La surface ROC est une variété continue de dimension 2 dans le cube unité de $\mathbb{R}^3$. On remarque que la surface ROC contient les courbes ROC de chacun des problèmes binaires (ϕ_1, ϕ_2), (ϕ_2, ϕ_3) et (ϕ_1, ϕ_3) qui sont obtenues en prenant les intersections de la surface ROC avec les plans orthogonaux à chaque axe du cube unité. Dans le cas où s n'a pas la capacité de discriminer entre les trois lois, *i.e.* quand $F_{s,1} = F_{s,2} = F_{s,3}$, la surface ROC revient à la surface délimité par le triangle qui connecte les points $(1, 0, 0)$, $(0, 1, 0)$ et $(0, 0, 1)$, on a alors $\mathrm{ROC}(\mathrm{s}, \alpha, \gamma) = 1 - \alpha - \gamma$. A l'opposé, dans le cas séparable (i.e. les supports des lois conditionnelles sont disjoints), la surface ROC optimal coïncide avec la surface du cube unité $[0, 1]^3$.

Pour garder le lien entre la surface ROC et ces sections, on introduit la notation suivante: $\forall \alpha \in [0, 1]$, $\mathrm{ROC}_{\phi_\mathrm{k}, \phi_{\mathrm{k}+1}}(\mathrm{s}, \alpha) = 1 - \mathrm{F}_{\mathrm{s,k}+1} \circ \mathrm{F}^{-1}_{\mathrm{s,k}}(1 - \alpha)$, où on a utilisé la définition suivante de l'inverse généralisée F: $F^{-1}(u) = \inf\{t \in]-\infty,\ +\infty] : F(t) \geq u\}$, $u \in [0, 1]$.

La surface ROC d'une fonction de scoring s peut être obtenue par le tracé de l'extention continue de la surface paramétrique suivante: $[0, 1]^2 \to \mathbb{R}^3, (\alpha, \gamma) \mapsto (\alpha, \mathrm{ROC}(\mathrm{s}, \alpha, \gamma), \gamma)$ où

$$\begin{aligned} \mathrm{ROC}(\mathrm{s}, \alpha, \gamma) &= \left(F_{s,2} \circ F^{-1}_{s,3}(1 - \gamma) - F_{s,2} \circ F^{-1}_{s,1}(\alpha)\right)_+ \\ &= \left(\mathrm{ROC}_{\phi_1, \phi_2}(\mathrm{s}, 1 - \alpha) - \mathrm{ROC}_{\phi_3, \phi_2}(\mathrm{s}, \gamma)\right)_+ \ , \end{aligned}$$

avec la notation $u_+ = \max(0,\ u)$, pour tout réel u.

D'autres notions de surface ROC ont été considérées dans la littérature, selon le problème d'apprentissage pris en compte et le but poursuivi. Dans le contexte de la reconnaissance de formes, ils fournissent un outil visuel de performance de classification, comme dans [Ferri *et al.*, 2003] (voir aussi [Fieldsend & Everson, 2005], [Fieldsend & Everson, 2006] et [Hand & Till, 2001]) du point de vue *one-versus-one* ou dans [Flach, 2004] quand on adopte l'approche *one-versus-all*. Le concept d'analyse ROC décrit ci-dessus est plus adapté au cas où un ordre naturel existe sur l'ensemble des classes, comme dans le cas de la régression ordinale, voir [Waegeman *et al.*, 2008c].

ROC-optimalité et fonctions de scoring optimales

La surface ROC fournit un outil visuel pour l'évaluation des performances de ranking d'une fonction de scoring. On a montré que l'optimalité pour les fonctions de scoring est équivalente à l'optimalité au sens de la surface ROC. On voit ainsi que la surface ROC fournit une caractérisation complète des performances en ranking d'une fonction de scoring pour le problème à 3 classes. De plus, optimiser la surface ROC revient à optimiser simultanément les courbes ROC reliées aux deux paires de lois (ϕ_1, ϕ_2) et (ϕ_2, ϕ_3).

Volume sous la surface ROC (VUS)

Dans le cas binaire, un critère standard de la performance de ranking est l'aire sous la courbe ROC (ou AUC). De la même manière, on peut considérer le *volume sous la surface* ROC (VUS en forme abrégée) dans le cadre à 3-classes. On suit ici [Scurfield, 1996] mais on mentionne que d'autres notions de surface ROC peuvent être trouvées dans la littérature, menant à d'autres critères, aussi appelés VUS, tel que celui introduit dans [Hand & Till, 2001]. On définit le VUS d'une fonction de scoring réelle s par :

$$\mathrm{VUS(s)} = \int_0^1 \int_0^1 \mathrm{ROC(s}, \alpha, \gamma) \, \mathrm{d}\alpha \mathrm{d}\gamma \ .$$

VUS-optimalité

On considère maintenant l'optimalité par rapport au critère du VUS et on fournit des expressions du déficit de VUS pour toute fonction de scoring qui éclairent le lien avec les maximisers d'AUC pour les sous-problèmes binaires. Sous l'hypothèse MLR, on a, pour toute fonction de scoring s et pour toute fonction optimale de scoring s^*: $\mathrm{VUS(s)} \leq \mathrm{VUS(s^*)}$. On note la valeur maximale du $\mathrm{VUS}^* = \mathrm{VUS(s^*)}$. Ce résultat montre que les fonctions optimales de scoring coïncident avec les éléments optimaux au sens du VUS. Cette assertion justifie l'utilisation de stratégies basées sur la maximisation du VUS empirique pour le problème du ranking à K-classes.

Quand l'hypothèse MLR n'est pas remplie, le VUS peut aussi être utilisé comme critère de performance, autant dans le contexte de la classification multi-classes

([Landgrebe & Duin, 2006], [Ferri *et al.*, 2003]) que dans le cadre de la régression ordinale ([Waegeman *et al.*, 2008c]). Cependant, on ne peut pas dire que les maximiseurs du VUS sont optimaux pour les problèmes de ranking. Supposons que l'hypothèse MLR est vérifiée. Alors, pour toute fonction de scoring s et toute fonction optimale de scoring s^*, on a

$$\mathrm{VUS}(s^*) - \mathrm{VUS}(s) \leq \left(\mathrm{AUC}^*_{\phi_1,\phi_2} - \mathrm{AUC}_{\phi_1,\phi_2}(s)\right) + \left(\mathrm{AUC}^*_{\phi_2,\phi_3} - \mathrm{AUC}_{\phi_2,\phi_3}(s)\right).$$

Ce résultat est la pierre angulaire du chapitre 4, dans lequel le but est de trouver un consensus entre une bonne fonction pour le problème 1 contre 2 et une bonne fonction pour le problème 1 contre 3 créant ainsi une fonction adéquate pour le problème à 3 classes.

Résumé chapitre 2

Dans le chapitre 2, la fonction de scoring s est fixée et on s'intéresse à estimer et trouver des régions de confiance pour la surface ROC. La surface ROC dépend des lois conditionnelles de $s(X)$ sachant le label auxquelles on n'a pas accès. Ce qu'on possède est un ensemble de données $\mathcal{D}_n = \{(s(X_i), Y_i\}$ qui nous permet d'estimer la surface ROC. On a mis en place une procédure d'estimation non-paramétrique pour estimer la surface ROC et donner une approximation forte de cet estimateur. Ensuite , grâce à une procédure du bootstrap lissée, on a trouvé des régions de confiance pour cet estimateur.

Travaux antérieurs

Plusieurs auteurs ont abordés le problème d'estimation de la courbe ROC et de la construction de bandes de confiance. La première étude importante a été faite dans [Hsieh & Turnbull, 1996], où des estimateurs non-paramétriques et semi-paramétriques de la courbe ROC sont proposés et les auteurs ont montré la convergence asymptotique de ces estimateurs. Dans [Macskassy & Provost, 2004], plusieurs métriques sur les courbes ROC sont introduites utilisées pour construire des bandes de confiance en reliant un certain nombre d'intervalles de confiance. Dans [Hall *et al.*, 2004], les auteurs utilisent un estimateur non-paramétrique couplé à une procédure de bootstrap lissé pour créer des bandes de confiance et montrent la convergence ponctuelle de ces bandes de confiance. Des théorèmes de convergence uniforme pour l'estimateur non-paramétrique sont donnés dans [Clémençon *et al.*, 2008] et [Gu & Ghosal, 2008] ainsi que pour les probabilités de couverture. Dans ce chapitre, le but est d'obtenir des régions de confiance pour la surface ROC.

Estimation de la surface ROC

Comme dans [Li & Zhou, 2009], on a estimé la surface ROC en remplaçant les fonctions de réparation par leur version empirique en utilisant un ensemble de données

$\mathcal{D}_n = \{(Z_i, Y_i)\}_{1 \leq i \leq n}$ où les (Z_i, Y_i) sont des copies i.i.d de (Z, Y), ce qui nous donne

$$\forall (\alpha, \gamma) \in [0,1]^2, \widehat{\mathrm{ROC}}(Z, \alpha, \gamma) = \left(\hat{F}_2 \circ \hat{F}_3^{-1}(1-\gamma) - \hat{F}_2 \circ \hat{F}_1^{-1}(\alpha) \right)_+$$

$\hat{F}_k = \frac{1}{n_k} \sum_{i=1}^n \mathbb{I}\{Y_i = k\} \mathbb{I}\{(x - Z_i) \geq 0\}$ où $n_k = \sum_{i=1}^n \mathbb{I}\{Y_i = k\}$ est le nombre (aléatoire) d'observations dont le label est k dans l'échantillon.

Pour obtenir une version régulière de la fonction de réparation $\tilde{F}_k$, on choisit de remplacer la fonction indicatrice par une fonction régularisante $K_h : \mathbb{R}^d \to \mathbb{R}$ telle que $K_h(u) = h^{-1} K(h^{-1} u)$ où K est un noyau de Parzen-Rosenblatt non nul dans un voisinage de 0 et qui vérifie $\int K(v) dv = 1$. Le paramètre $h > 0$ est appelé la fenêtre de lissage.

Normalité asymptotique

Les surfaces ROC sont de nature fonctionnelle et on utilise la norme infini pour évaluer la distance entre deux surfaces ROC. On établit également une approximation forte du processus de fluctuation,

$$r_n(\alpha, \gamma) = \sqrt{n} \left(\widehat{\mathrm{ROC}}(\alpha, \gamma) - \mathrm{ROC}(\alpha, \gamma) \right), (\alpha, \gamma) \in [0,1]^2.$$

On utilise des hypothèses qui sont classiques dans la théorie de l'approximation forte (voir [Csorgo & Revesz, 1981]). Sous ces hypothèses, on montre deux résultats sur l'estimateur empirique de la surface ROC, le premier établissant la consistance forte (c'est à dire presque surement) de l'estimateur. Quant au second, il donne la normalité asymptotique de l'estimateur dans lequel la loi limite est exprimée en fonction de ponts browniens, plus précisément

$$\begin{aligned} &\frac{1}{\sqrt{p_2}} B_1^{(n)}(F_{s,2}(F_{s,3}^{-1}(1-\gamma))) + \frac{1}{\sqrt{p_3}} \frac{f_{s,2}(F_{s,3}^{-1}(1-\gamma))}{f_{s,3}(F_{s,3}^{-1}(1-\gamma))} B_2^{(n)}(\gamma) \\ &\qquad - \frac{1}{\sqrt{p_2}} B_1^{(n)}(F_{s,2}(F_{s,1}^{-1}(\alpha))) + \frac{1}{\sqrt{p_1}} \frac{f_{s,2}(F_{s,1}^{-1}(\alpha))}{f_{s,1}(F_{s,1}^{-1}(\alpha))} B_3^{(n)}(\alpha), \quad (1) \end{aligned}$$

où les $B_j^{(n)}$ sont des ponts browniens indépendants. La vitesse de convergence est un $O\left(\frac{\ln(n)}{\sqrt{n}}\right)$ ce qui est classique pour un théorème central limite fonctionnel d'approximation forte. Dans [Li & Zhou, 2009], un théorème asymptotique similaire a été prouvé mais l'approximation est valable seulement en loi.

Bootstrap lissé

Dans cette partie, on veut construire des régions de confiance pour la surface ROC en utilisant l'approche bootstrap introduite par [Efron, 1979]. Le but est d'étendre les résultats établis dans le cas binaire pour la courbe ROC dans [Bertail *et al.*, 2008].

Ce dernier suggère de considérer, comme estimation de la loi du processus de fluctuation $r_n = \{r_n(\alpha,\gamma)\}_{(\alpha,\gamma)\in[0,1]^2}$, la loi conditionnelle sachant $\mathcal{D}_n$ du processus bootstrap de fluctuation

$$r_n^*(\alpha,\gamma) = \sqrt{n}(\mathrm{ROC}^*(\alpha,\gamma) - \widehat{\mathrm{ROC}}(\alpha,\gamma)), ((\alpha,\gamma)) \in [0,1]^2$$

où ROC^* est la surface ROC correspondant à l'échantillon $\mathcal{D}_n^* = \{(Z_i^*, Y_i^*)\}_{1\leq i\leq n}$ de paires aléatoires i.i.d. de loi $\tilde{P}_n$ proche de $\mathcal{P}_n$. Le choix naïf de $\tilde{P}_n = \hat{P}_n$, la loi empirique, n'est pas le meilleur car l'estimation de la surface ROC implique le processus quantile. Dans cette situation, le bootstrap lissé, qui consiste à prendre une version régularisée des fonctions de répartition empirique, améliore le bootstrap naïf d'un point de vue théorique et pratique.

La procédure de construction de régions de confiance pour la surface ROC grâce au bootstrap lissé est la suivante. Premièrement, à partir d'un échantillon $\mathcal{D}_n$, on calcule la surface ROC empirique. Ensuite on tire un échantillon $\mathcal{D}_n^*$ à partir de la loi lissée

$$P(dz,y) = \frac{n_1}{n}\mathbb{I}\{Y=1\}\tilde{F}_1(dz) + \frac{n_2}{n}\mathbb{I}\{Y=2\}\tilde{F}_2(dz) + \frac{n_3}{n}\mathbb{I}\{Y=3\}\tilde{F}_3(dz).$$

En utilisant ce jeu de données, on calcule la version bootstrap des fonctions de répartition empirique pour chacune des classes F_1^*, F_2^*, F_3^* à partir de $\mathcal{D}_n^*$. Il fournit des régions de confiance au niveau $1-\varepsilon$ dans l'espace ROC à partir de l'échantillon $\mathcal{D}_n = \{Z_i, Y_i\}$.

On étudie les propriétés asymptotiques de cette procédure de bootstrap. Il est important de remarquer que le résultat est de nature fonctionnelle car dans les applications l'estimation de la surface ROC, ou au moins une partie de celle-ci, est ce que l'on recherche. En utilisant les même hypothèses que pour le théorème centrale limite fonctionnel précédent et que les versions régularisée des fonctions de répartition $\tilde{F}_1, \tilde{F}_2, \tilde{F}_3$ sont calculées avec le noyau $K_{h_n}(u)$ avec $h_n \downarrow 0$ quand $n\to\infty$ de façon à ce qu'on ait $nh_n^3 \to \infty$ et $nh_n^5 \log^2(n) \to 0$. Alors, la sortie de l'algorithme de bootstrap lissé est telle que:

$$\sup_{t\in\mathbb{R}} |P^*(\|r_n^*\|_\infty \leq t) - P(\|r_n\|_\infty \leq t)| = O_{\mathbb{P}}\left(\frac{log(h_n^{-1})}{\sqrt{nh_n}}\right)$$

En prenant la fenêtre h_n de l'ordre $1/\log(n^{2+\varepsilon})n^{1/5}$ avec $\varepsilon > 0$ cela nous donne une erreur d'approximation de l'ordre de $\left(n^{-2/5}\right)$ à un facteur logarithme près pour l'estimation de la loi bootstrap. Cette vitesse de convergence est plus lente que celle de l'approximation gaussienne donnée dans le théorème précédent, l'algorithme bootstrap pour la surface ROC est très intéressant du point de vue computationnel. Notamment, elle évite d'avoir à estimer les densités $f_{s,1}$, $f_{s,2}$ et $f_{s,3}$. De plus, la vitesse atteinte par le bootstrap lissé est bien plus rapide que celle du bootstrap naïf qui est d'ordre $\left(n^{-1/4}\right)$.

Résumé du chapitre 3

Dans le chapitre 3, le but est d'étendre le principe de minimisation du risque empirique, d'un point de vue pratique, à la situation où l'estimateur du risque prend la forme d'une U-statistique, ce qui est le cas du VUS. Dans ce cas, le calcul de l'estimateur empirique est est difficilement faisable dès que le nombre de données est grand car il implique une moyenne avec $O(n^{d_1+\ldots+d_K})$ termes, quand on considère une U-statistique, de degrés $(d_1, \ldots, d_K)$. On se propose d'étudier une version Monte-Carlo du risque empirique basée sur seulement $O(n)$ termes qui peut être vu comme une U-statistique incomplète. Cette technique de tirage avec remise a été proposé par [Blom, 1976] dans le contexte de l'estimation ponctuelle et elle permet de garder les propriétés de réduction de la variance tout en préservant les vitesses d'apprentissages.

Motivation

Dans le cas du ranking multipartie, le critère principal est le VUS et sa contrepartie empirique prend la forme d'une U-statistique. Pour l'évaluer, on a besoin de K échantillons indépendants un par classe et on appelle n_k la taille de l'échantillon de la classe k. Dans ce cas, le VUS d'une fonction sans ex-æquo s'écrit

$$\widehat{\mathrm{VUS}}_{\mathbf{n}}(s) = \frac{\sum_{i_1=1}^{n_1} \cdots \sum_{i_K=1}^{n_K} \mathbb{I}\left\{s(X_{i_1}^{(1)}) < \ldots < s(X_{i_K}^{(K)})\right\}}{n_1 \times \cdots \times n_K}. \tag{2}$$

On voit facilement que cette somme implique $n_1 \times \cdots \times n_K$. Bien qu'une méthode astucieuse permet de calculer le VUS en $O(n \ln n)$ itérations, trouver le maximiseur empirique du VUS a pour complexité $n_1 \times \cdots \times n_K$. Or ce nombre est prohibitif dans les cas venus du web où les tailles de chaque classe sont de l'ordre de 10^6.

Approximation uniforme des U-statistiques généralisées

Si on se donne K échantillons $(X_1^{(k)}, \ldots, X_{n_k}^{(k)})$, $1 \leq k \leq K$, la U-statistique de degrés $(d_1, \ldots, d_K)$ associée au noyau $H : \mathcal{X}_1^{d_1} \times \cdots \times \mathcal{X}_K^{d_K} \to \mathbb{R}$ prend la forme suivante,

$$U_{\mathbf{n}}(H) = \frac{\sum_{I_1} \cdots \sum_{I_K} H(\mathbf{X}_{I_1}^{(1)}; \mathbf{X}_{I_2}^{(2)}; \ldots; \mathbf{X}_{I_K}^{(K)})}{\binom{n_1}{d_1} \times \cdots \times \binom{n_K}{d_K}}. \tag{3}$$

La principale propriété de cette statistique est que $U_{\mathbf{n}}(H)$ a la variance minimum parmi les estimateurs non biaisés de $\theta(H) = \mathbb{E}[H(X_1^{(1)}, \ldots, X_{d_1}^{(1)}, \ldots, X_1^{(K)}, \ldots, X_{d_k}^{(K)})]$. Pour étudier le comportement asymptotique de cet estimateur, l'hypothèse classique est de supposer que les rapports n_k/n tendent vers des constantes λ_k quand n tend vers ∞. En utilisant des techniques de linéarisation (voir [Hoeffding, 1948]), on peut montrer que $U_{\mathbf{n}}(H)$ tend vers $\theta(H)$ à la vitesse $1/\sqrt{n}$.

Quand les tailles des échantillons sont grandes, la U-statistique n'est pas calculable et on l'estime par la quantité suivante

$$\widetilde{U}_B(H) = \frac{1}{B} \sum_{(I_1, \ldots, I_K) \in \mathcal{D}_B} H(X_{I_1}^{(1)}, \ldots, X_{I_K}^{(K)}), \tag{4}$$

où $\mathcal{D}_B$ est un ensemble créé grâce à un tirage avec remise. En pratique, cette somme doit comporter $O(n)$ termes pour surmonter les difficultés computationnelles. Cet estimateur est non biaisé mais sa variance est plus grande que celle de $U_{\mathbf{n}}(H)$ et vaut $\mathrm{Var}(\widetilde{U}_B(H)) = (1 - 1/B)\mathrm{Var}(U_{\mathbf{n}}(H)) + O(1/B)$.

Le résultat principal de ce chapitre est le suivant. Pour des noyaux H appartenant à des classes $\mathcal{H}$ de dimension de Vapnik Chervonenkis(VC) finie V, on a obtenu des inégalités maximales de déviation pour la quantité $\sup_{H \in \mathcal{H}} \left| \widetilde{U}_B(H) - U_{\mathbf{n}}(H) \right|$ de l'ordre de

$$2\sqrt{\frac{2\mathcal{V}\log(1+\kappa)}{\kappa}} + \sqrt{\frac{\log(2/\delta)}{\kappa}} + \sqrt{\frac{\mathcal{V}\log(1+\#\Lambda) + \log(4/\delta)}{B}},$$

où $\kappa = \min\{\lfloor n_1/d_1 \rfloor, \ldots, \lfloor n_K/d_K \rfloor\}$.
Dans le cas du ranking multipartite, ces inégalités impliquent que si la fonction de scoring appartient à une classe de VC dimension V alors on a, avec probabilité $1-\delta$, $max_{s \in \mathcal{S}} \mathrm{VUS(s)} - \mathrm{VUS}(\widehat{\mathrm{s}}_{\mathrm{B}})$ qui est majoré par $\sqrt{\frac{\mathcal{V}\log(\#\Lambda/\delta)}{n}}$ où $\#\Lambda = n_1 \times \ldots \times n_K$. De plus, on a montré que dans des cas simulés et réels, les même conclusions sont observés, c'est à dire que l'estimation du VUS avec très peu de données est presque aussi bonne qu'avec toutes les données. Dans certains cas, on a gagné un facteur 1000 dans le calcul de la fonction de scoring d'apprentissage.

Résumé des chapitres 4, 5 et 6

Résumé du chapitre 4

Dans le chapitre 1, le but du ranking multipartite a été défini de manière quantitative, on se tourne vers la réduction de ce problème d'apprentissage à une série de tâches de ranking binaires, avec un mode de fonctionnement similaire à la méthode de comparaison par paires, aussi connu sous le nom de méthode "all versus all" en reconnaissance de formes multi-classes. L'approche de l'ordonnancement à K classes développée dans cet article consiste à le voir comme une superposition de tâches d'ordonnancement binaire. En effet, les solutions sont les règles d'ordonnancement qui sont à la fois optimales pour les $K-1$ des problèmes d'ordonnancement binaire relativement à toutes les paires possibles d'étiquettes consécutives (*ie* paires de lois de classe consécutives). Du coup, la procédure que nous proposons ici est implémentée en deux étapes. La première consistant à résoudre les sous problèmes d'ordonnancement binaire séparément, construisant ainsi une collection de règles de scoring. La seconde étape est le calcul d'une fonction de scoring médiane, reliée à la collection obtenue à la première étape et basée sur une notion de distance entre les règles de scoring, la dissimilarité étant mesurée par le τ de Kendall. Il a été montré qu'une telle médiane existe toujours dans le cas où les fonctions de scoring que l'on cherche à agréger sont constantes par morceaux, et son calcul est possible. On établit ensuite que le consensus résultant est une règle de ranking consistante, à condition que la méthode de ranking utilisée pour résoudre les sous-problèmes de ranking binaires soit elle-même consistante.

Agrégation par paire: du ranking binaire au ranking à K-classes

Dans cette partie, on propose une stratégie pratique pour construire des fonctions de scoring qui approximent les fonctions de scoring optimales pour le ranking à K-classes à partir d'observations labellisées. Le principe de cette stratégie est d'agréger des fonctions de scoring obtenues sur les sous-problèmes binaires. On insiste sur le fait que cette situation est très différente de celle de la classification multi-classes où l'agrégation se résume à prendre une combinaison linéaire, ou un vote majoritaire, des classifieurs binaires (pour les approches "*one against one*" et "*one versus all*", on se réfère à [Allwein *et al.*, 2001], [Hastie & Tibshirani, 1998], [Venkatesan & Amit, 1999], [Debnath *et al.*, 2004], [Dietterich & Bakiri, 1995], [Beygelzimer *et al.*, 2005b], [Beygelzimer *et al.*, 2005a]). On propose ici, dans le cas du ranking à K-classes, une approche barycentrique basée sur une métrique pour construire une fonction de scoring agrégée à partir de la collection des fonctions de scoring estimées sur chaque sous-problème binaire.

Fonctions de scoring médiane et agrégation optimale

Toute fonction de scoring induit une relation d'ordre sur l'espace d'entrée $\mathbb{R}^d$ et, pour le problème de ranking considéré ici, une mesure de similarité entre deux fonctions de scoring doit uniquement prendre en compte la ressemblance entre les ordres induits par chacune des fonctions. Ici, on propose une mesure d'entente entre les fonctions de scoring qui est basée sur la version probabiliste du τ de Kendall pour une paire de variables aléatoires. Cette mesure d'accord entre les fonctions de scoring s_1 ets_2 coïncide en effet avec le τ de Kendall entre les variables aléatoires $s_1(X)$ et $s_2(X)$. Notons que la contribution des deux derniers termes dans la définition de $\tau(s_1, s_2)$ sont strictement positifs aux endroits où les fonctions de scoring sont constantes par morceaux. Les fonctions de scoring constantes par morceaux ont une place spéciale pour le problème du ranking puisque il existe des méthodes récursives de partitionnement de l'espace d'entrée basée sur des arbres (voir [Clémençon & Vayatis, 2009b]).

On peut maintenant définir la notion de fonction de scoring médiane qui fait le consensus de plusieurs fonctions de scoring réelles $\Sigma_K = \{s_1, \ldots, s_{K-1}\}$ sur une classe donnée de fonctions candidates $\mathcal{S}_1$. Une fonction de scoring médiane $\overline{s}$ pour $(\mathcal{S}_1, \Sigma_K)$ vérifie:

$$\sum_{k=1}^{K-1} \tau(\overline{s}, s_k) = \sup_{s \in \mathcal{S}_1} \sum_{k=1}^{K-1} \tau(s, s_k).$$

On peut montrer qu'une fonction de scoring médiane, basée sur $K-1$ fonctions de scoring optimales sont aussi des solutions optimales pour le problème global de ranking à K-classes. Cela suggère d'implémenter la procédure en deux étapes suivante, qui consiste en 1) à résoudre les sous-problèmes binaire de ranking reliés à chaque paire consécutive $(k, k+1)$ de labels, produisant une fonction de scoring s_k, pour $1 \leq k < K$, et 2) à calculer une fonction médiane selon la définition précédente, quand c'est réalisable, sur une classe $\mathcal{S}_1$ de fonctions de scoring. Au-delà de la difficulté à résoudre chacun des sous-problèmes de ranking séparément (pour le moment, on fait référence à [Clémençon & Vayatis, 2009b] pour une discussion sur la nature du problème de ranking binaire), la performance/complexité de la méthode esquissée ci-dessus est réglée par la richesse de la classe $\mathcal{S}_1$ de fonctions de scoring candidates: des classes trop complexes rendent le calcul de la médiane impossible, alors que des classes trop simples peuvent ne pas contenir des fonctions de scoring suffisamment performantes.

Consistance de l'agrégation par paire et autre stratégies pour le ranking à K-classes

Dans cette partie, on suppose qu'un ensemble de données $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ est disponible et composé de n copies i.i.d. de la paire (X, Y). Notre but est d'apprendre à partir de l'échantillon $\mathcal{D}_n$ comment construire une fonction de scoring $\widehat{s}_n$ telle que sa surface ROC est aussi proche

que possible de la surface ROC optimale. On propose de considérer une forme faible de consistance qui se base sur le VUS. On dit que la suite $\{s_n\}$ est dite VUS-consistante si $\mathrm{VUS}^* - \mathrm{VUS}(s_n) \to 0$ en probabilité. Pour établir le résultat principal, on a besoin d'une hypothèse supplémentaire sur la distribution du couple (X, Y) dite de marge. Dans la littérature d'apprentissage statistique, la condition est dite condition de bruit et remonte au travail de Tsybakov [Tsybakov, 2004]. Elle a été adaptée dans le cadre du ranking binaire dans [Clémençon *et al.*, 2008]. Elle est également utilisée dans les chapitres 5 et 7. Dans le chapitre 7, les conditions de bruit faible sont largement étudiées et reliées entre elles.

On a également besoin d'utiliser la notion de AUC-consistance pour les sous-problèmes de ranking binaire. Pour k dans $\{1, \ldots, K-1\}$, une suite $(s_n)_{n\geq 1}$ de fonctions de scoring est dite AUC-consistante pour le problème de ranking binaire (ϕ_k, ϕ_{k+1}) si elle vérifie : $\mathrm{AUC}_{\phi_k,\phi_{k+1}}(s_n) \to \mathrm{AUC}^*_{\phi_k,\phi_{k+1}}$ en probabilité.

Le résultat principal de consistance de ce chapitre qui concerne la procédure d'agrégation via le τ de Kendall qui est décrite ci-dessus. On montre que la notion de fonction de scoring médiane introduite précédemment préserve l'AUC-consistance pour les sous-problèmes de ranking binaire et produit ainsi une fonction de scoring VUS-consistante pour le problème de ranking à K-classes.

Résumé du chapitre 5

Le but du chapitre 5 est de construire une fonction de scoring constante par morceaux telle que sa surface ROC associée soit proche de la surface ROC optimale. L'algorithme proposé est une version empirique d'un schéma d'approximation de la surface ROC optimale par une fonction affine par morceaux.

Cet algorithme s'inspire de la procédure TreeRank proposée dans le cas binaire, voir [Clémençon & Vayatis, 2009b]. Cet algorithme top-down consiste à créer un arbre de décision par récurrence en résolvant à chaque étape un problème de classification binaire pondérée. La difficulté principale pour passer au cas multipartite vient du fait que la règle optimale de séparation des données dans une feuille ne peut pas s'écrire comme un problème de classification pondérée. Pour surmonter cette difficulté, on se propose de mettre en compétition les règles de décision trouvées à partir de sous-problèmes binaires. Cette procédure est appelée TreeRank Tournament et on montre que sa version probabiliste nous donne bien un schéma d'approximation pour la surface ROC optimale. De plus, la version empirique est consistante sous de bonnes conditions. On a également obtenu une vitesse de convergence très lente à cause de l'accumulation des erreurs à chaque pas de la récurrence.

Fonctions constantes par morceaux

Pour une partition $\mathcal{C}_N = (C_l)_{1\leq l\leq N}$, la fonction de scoring associée est $\forall x \in \mathcal{X}, s_N(x) = \sum_{l=1}^N a_j \mathbb{I}\{x \in C_l\}$, où $\{a_1, \ldots, a_N\}$ sont les valeurs prises par la fonction de scoring. Dans le cas $K = 3$, la surface ROC associée à s_N est composée de N^2

morceaux de plan et la surface ne dépend pas des valeurs des a_l mais de leur ordre. Quitte à réordonner les C_l on peut donc supposer que les a_l sont rangés par ordre décroissant. En utilisant une décomposition en éléments finis, on peut exprimer la surface ROC associée à s_N ainsi que le volume sous la surface ROC.

Rappel sur le cas binaire

Dans [Clémençon & Vayatis, 2009b] (voir aussi [Clémençon *et al.*, 2011b]), l'algorithme de ranking TreeRank optimisant directement la courbe ROC de façon récursive a été proposé. Il produit une partition orientée de l'espace $\mathcal{X}$ définissant ainsi un ranking pour lequel tous les éléments d'une même cellule sont égaux. La forme finale est un arbre binaire orienté de gauche à droite de profondeur $J \geq 1$. La racine représente tout l'espace $\mathcal{X}$ et chaque nœud (j,k), $j < J$ et $k \in \{0, \ldots, 2^j - 1\}$ correspond à une sous-espace $\mathcal{C}_{j,k} \subset \mathcal{X}$ dont les enfants $\mathcal{C}_{j+1,2k}$ et $\mathcal{C}_{j+1,2k+1}$ sont des ensembles disjoints tels que $\mathcal{C}_{j,k} = \mathcal{C}_{j+1,2k} \cup \mathcal{C}_{j+1,2k+1}$. Au départ la fonction de scoring vaut $s_1(x) = \mathbb{I}\{x \in \mathcal{C}_{0,0}\} \equiv 1$ et à la profondeur J, $s_J(x) = \sum_{l=0}^{2^J-1}(2^J - l) \cdot \mathbb{I}\{x \in \mathcal{C}_{J,l}\}$. Pour partager chaque cellule, l'algorithme résout un problème de classification pondéré que revient à maximiser l'AUC empirique, voir [Clémençon *et al.*, 2011a] pour plus de détails. Si on connaissait les lois conditionnelles de chaque classe, la version théorique de cet algorithme (où on résout parfaitement le problème de maximisation de l'AUC théorique) est un schéma d'approximation du type éléments finis de la courbe ROC. En particulier, chaque point ajouté entre deux points existants $\alpha_{j,k}$ et $\alpha_{j,k+1}$ a sa tangente égale à $(\mathrm{ROC}^*(\alpha_{\mathrm{j,k+1}}) - \mathrm{ROC}^*(\alpha_{\mathrm{j,k}})/(\alpha_{\mathrm{j,k+1}} - \alpha_{\mathrm{j,k}})$. Le but du chapitre 5 est de transposer cette stratégie dans le cas multipartite.

Approximation de la surface ROC

Pour la procédure d'approximation, on se restreint au cas $K = 3$. La procédure récursive fonctionne de la façon suivante. Au départ, on a $C_{0,0} = \mathcal{X}$. Pour passer de la profondeur j à $j+1$ utilise la procédure suivante : pour chaque $C^*_{j,l}$ avec $l \in \{0, \ldots, 2^j - 1\}$ on résout les sous-problèmes binaires de partitionnement pour les problèmes k vs $(k+1)$ pour $k \in \{1, 2\}$. Cela nous fournit deux sous-espaces candidats : $\mathcal{C}^{(1)}_{j+1,2l}$ et $\mathcal{C}^{(2)}_{j+1,2l}$. On choisit celui qui maximise le critère du VUS dans la sous-partie $\mathcal{C}_{j,l}$ i.e. le critère

$$\begin{aligned}\mathrm{VUS}_{\mathcal{C}^*_{\mathrm{j,k}}}(\mathcal{C}^{(\mathrm{l})}_{\mathrm{j+1,2k}}) = {} & \mathrm{F}_3(\mathcal{C}^{(\mathrm{l})}_{\mathrm{j+1,2k}})(\mathrm{F}_1(\mathcal{C}^*_{\mathrm{j,k}}) - \mathrm{F}_1(\mathcal{C}^{(\mathrm{l})}_{\mathrm{j+1,2k}}))/2 \\ & + F_1(\mathcal{C}^{(l)}_{j+1,2k})F_2(\mathcal{C}^{(l)}_{j+1,2k})F_3(\mathcal{C}^{(l)}_{j+1,2k})/6 \\ & + (F_1(\mathcal{C}^*_{j,k}) - F_1(\mathcal{C}^{(l)}_{j+1,2k}))(F_2(\mathcal{C}^*_{j,k}) - F_2(\mathcal{C}^{(l)}_{j+1,2k}))(F_3(\mathcal{C}^*_{j,k}) - F_3(\mathcal{C}^{(l)}_{j+1,2k}))/6.\end{aligned}$$

Cela nous donne $C^*_{j+1,2l}$ et $C^*_{j+1,2l+1} = C^*_{j,l} \setminus C^*_{j+1,2l}$. Cette procédure fournit une fonction de scoring constante par morceaux et on montre que la surface ROC associée converge vers la surface ROC optimale à une vitesse classique en théorie de

l'approximation c'est à dire $O(N^{-2})$ où N est le nombre de feuilles terminales. Les hypothèses nécessaires pour prouver ce résultat sont des conditions de régularité sur les lois conditionnelles de chaque classe. En particulier, on n'autorise pas la surface ROC optimale à avoir des tangentes verticales.

L'algorithme TreeRank Tournament

On a accès à un jeu de données $\mathcal{D}_n = \{(X_i, Y_i), i = 1, \ldots, n\}$ et on veut construire une fonction de scoring qui partitionne et ordonne l'espace $\mathcal{X}$. L'algorithme TreeRank Tournament est la version empirique de la procédure précédente i.e. on utilise le jeu de données pour résoudre les problèmes de classification binaire ainsi que pour déterminer le vainqueur du tournoi à chaque étape. Notons en particulier que dans la pratique le nombre de classes n'est pas limité et que la décomposition en sous-problèmes binaires n'est pas restreinte à k vs $(k+1)$. Les résultats pratiques de cet algorithme sont présentés dans le chapitre 6.

Pour cet algorithme, on obtient des garanties théoriques qui sont la consistance ainsi que la vitesse de convergence en norme L_1. Les conditions pour prouver la consistance sont les mêmes que pour la convergence du schéma d'approximation. Par contre pour obtenir la vitesse on a besoin d'une hypothèse de marge, essentiellement pour prouver que les cellules de la partition empirique sont proches des cellules du schéma d'approximation. La vitesse obtenue est en $O(1/n^{\frac{a^J}{2(1+a)^J}})$ et on remarque que cette vitesse se dégrade avec quand la profondeur augmente. C'est du à l'empilement des erreurs d'approximation des cellules optimales à travers la procédure de récurrence.

Résumé du chapitre 6

Le chapitre 6 étudie les performances numériques des algorithmes de ranking présentés dans les chapitres 4 et 5. Plusieurs jeux de données sont considérés pour comprendre comment les algorithmes réagissent aux différentes situations. On simule des jeux données avec des lois gaussiennes ou uniformes respectant l'hypothèse de monotonie des rapports de vraisemblance. On a aussi crée un jeu de données dans lequel les supports des lois conditionnelles ne sont pas les mêmes. On utilise également des jeux de données réels disponible sur internet. En particulier, on a fait plusieurs expériences sur ces jeux de données en supprimant des classes trop peu représentées pour voir l'influence des probabilités a priori i.e. les p_k.

Pour évaluer les résultats, on a évalué le VUS à travers une procédure de validation croisée. Précisément, pour un jeu de données fixé, on l'a partitionné en 5 et on a utilisé 4 parts pour calculer la fonction de scoring puis on a évalué le VUS empirique grâce à la dernière part. On fait ça pour chaque part et on fait la moyenne pour trouver un estimateur du VUS. Pour plus de stabilité on a répété cette procédure 5 fois et on a fait la moyenne. Pour évaluer la stabilité de l'estimation du VUS on calcule également l'écart-type des 25 évaluations

du VUS empirique. On a également considéré un autre critère, le VUS local : $\mathrm{LocVUS}(s, u_0) = \mathbb{P}\{s(X) < s(X') < s(X''), s(X'') > Q(s, u_0) | Y = 1, Y' = 2, Y'' = 3\}$ qui est la généralisation au cas [Clémençon & Vayatis, 2007] de l'AUC locale présentée dans [Clémençon & Vayatis, 2007]. Ce critère permet de comparer la qualité des fonctions de scoring pour les grandes valeurs. Ce genre de critère est particulièrement important pour les cas venus de fouilles de données où l'on souhaite retrouver les meilleurs documents pour une requête donnée. Dans un deuxième temps, on compare les algorithmes introduits dans cette thèse à plusieurs algorithmes existants dans la littérature (RankBoost ([Xu & Li, 2007]), RankSVM ([Joachims, 2002]) et RankRLS [Pahikkala *et al.*, 2007]). L'algorithme RankBoost consiste à agréger des fonctions de scoring faibles en utilisant une perte exponentielle. Dans nos expériences, on a utilisé fonctions de seuillage comme fonctions de décisions. Les algorithmes RankSVM et RankRLS sont tous les deux des algorithmes à noyaux qui visent à minimiser des heuristiques de la forme

$$\frac{2}{n(n-1)} \sum_{i<j} d(Y_i - Y_j, s(X_i) - s(X_j)) + \lambda \|s\|_{\mathcal{K}}.$$

Dans le cas de RankSVM la perte hinge est utilisée alors que RankRLS se base sur la perte quadratique.

Parmi nos algorithmes, on peut voir une première tendance se dessiner. Tout d'abord, la décomposition en sous-problèmes binaires joue un rôle essentiel pour la procédure d'agrégation de fonctions de scoring. En particulier, quand les supports des lois conditionnelles ne sont pas les même, les décompositions où chaque problème contient toutes les classes (par exemple, les classes plus petites que k contre les classes plus grandes que k) sont bien meilleurs que les décompositions du type "classe k" vs "classe l". Pour certains jeux de données, les algorithmes qui sont proposés dans cette thèse sont bien meilleurs que les algorithmes de l'état de l'art. Les particularités de ces jeux de données sont le grand nombre d'observations mais surtout que la dimension de l'espace des observations est bien plus grande que le nombre de classes. On explique ce phénomène, en voyant nos algorithmes comme une réduction de la dimension sans perte de l'information contenue dans les données.

Résumé chapitre 7

Introduction

Dans le chapitre 7, on retourne au cas binaire et on s'intéresse aux vitesses minimax dans le cas du ranking. Pour cela, on regarde les estimateurs de type plug-in, c'est à dire qui estiment la fonction de régression et que l'on remplace dans la règle de ranking. Les classes de probabilités considérées sont des classes vérifiant plusieurs hypothèses sur la fonction de régression ainsi que sur la loi marginale des observations. Enfin, on met en place une procédure d'agrégation par poids exponentiels pour obtenir des règles de ranking adaptatives aux paramètres des classes de probabilités.

Contexte théorique

Ici, on introduit les hypothèses principales associées à la formulation du problème de ranking binaire et on rappelle les résultats importants qui sont utilisés dans l'analyse suivante, donnant ainsi une idée de la nature du problème du ranking.

Cadre probabiliste et première notations

Dans cette partie, (X, Y) est un couple de v.a., à valeurs dans l'espace produit $\mathcal{X} \times \{-1, +1\}$ où $\mathcal{X}$ est un sous-ensemble de $\mathbb{R}^d$,$d \geq 1$. La v.a. X est vue comme une observation aléatoire qui permet de prédire le label Y. Soit $p = \mathbb{P}\{Y = +1\}$ le taux d'observations positives. On note P la loi jointe de (X, Y), μ la marginale de X et $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$ la probabilité a posteriori, $x \in \mathcal{X}$. De plus, on suppose que la v.a. est continue.

Ranking binaire

Considérer un critère de performance prenant ses valeurs dans un espace fonctionnel conduit naturellement à de grandes difficultés en ce qui concerne l'analyse mathématique et l'implémentation d'algorithmes. Plusieurs auteurs ont traité le problème du ranking d'un point de vue *classification par paire*, [Freund *et al.*, 2003, Agarwal *et al.*, 2005, Clémençon *et al.*, 2005]. Dans ce cadre, l'objectif est de déterminer, étant donné deux paires indépendantes (X, Y) et (X', Y') de loi P, si $Y' > Y$ ou le contraire. Une règle de ranking est une fonction (mesurable) $r : \mathcal{X}^2 \to \{-1, +1\}$ telle que $r(x, x') = 1$ quand x est classée au dessus de x: plus une règle de ranking r est pertinente, plus la probabilité de mal ordonner deux observations indépendantes est petite. Formellement, les règles de ranking optimales sont celles que minimisent le risque de ranking $L(r) \stackrel{def}{=} \mathbb{P}\{r(X, X') \cdot (Y' - Y) < 0\}$.

Optimalité. Pour le critère de performances ci-dessus, la règle $r^*(x,x') = 2 \cdot \mathbb{I}_{\{\eta(x')>\eta(x)\}} - 1$ définie par la fonction de régression $\eta(x)$ est optimale. De plus, on peut exprimer de façon probabiliste l'excès de risque de ranking $\mathcal{E}(r) = L(r) - L^*$, avec $L^* = L(r^*)$. Pour toute règle de ranking r, on a: $\mathcal{E}(r) = \mathbb{E}\left[|\eta(X) - \eta(X')| \, \mathbb{I}\{r(X,X')(\eta(X') - \eta(X)) < 0\}\right]$.

La précision d'une règle de ranking est ici caractérisée par l'excès de risque de ranking $\mathcal{E}(r)$, le défi du point de vue statistique étant de construire une règle de ranking, basée sur un échantillon d'apprentissage (X_1,Y_1), $\ldots$, (X_n,Y_n) de paires i.i.d de loi P, avec asymptotiquement un petit excès de risque de ranking quand n grandit.

Fonction de ranking plug-in. Etant donnée la forme de la règle de ranking de Bayes r^*, il est naturel de considérer les règles de ranking dites *plug-in*, c'est à dire les règles de ranking obtenues en remplaçant la fonction η par un estimateur non-paramétrique $\widehat{\eta}_n(x)$, basé sur l'ensemble de données (X_1,Y_1), $\ldots$, (X_n,Y_n), dans l'équation (7.2): $\widehat{r}_n(x,x') \stackrel{def}{=} r_{\widehat{\eta}_n}(x,x')$, $(x,x') \in \mathcal{X}^2$.

Convexification du risque de ranking. D'un point de vue pratique, optimiser le risque de ranking est un vrai problème car la perte est non convexe. En classification, les substituts convexes ont été largement utilisés pour des raisons pratiques et également pour résoudre des problèmes théoriques ([Bartlett *et al.*, 2006], [Zhang, 2004] et [Lecué, 2006] par exemple). Ici, on propose de convexifier la perte par paire et d'utiliser cette perte de convexe pour agréger les règles de ranking (voir 7.3). Notons que la minimisation d'une perte par paire convexifiée a été étudiée dans [Clémençon *et al.*, 2008]. On appelle une fonction mesure $f : \mathcal{X} \times \mathcal{X}' \to [-1,1]$ règle de décision et on pose la v.a. $Z = (Y - Y')/2$. Muni de ces notations, on introduit la convexification du risque de ranking que l'on utilise dans ce chapitre. Pour toute fonction de décision f, le hinge risque de ranking est défini par $A(f) \stackrel{def}{=} \mathbb{E}\phi\left(-f(X,X') \cdot Z\right)$. où $\phi(x) = \max(0, 1+x)$.

Hypothèses supplémentaires

Les règles de ranking optimales peuvent être définies comme celles ayant la meilleure vitesse de convergence de $\mathcal{E}(\widehat{r}_n)$ vers 0, quand $n \to +\infty$. Ainsi, cette vitesse dépend de la loi P. En suivant les traces de [Audibert & A.Tsybakov, 2007], on adopte le point de vue *minimax*, qui consiste à considérer une classe $\mathcal{P}$ de loi P et de dire que $\widehat{r}_n$ est optimale si elle atteint la meilleure vitesse de convergence sur cette classe:

$$\sup_{P \in \mathcal{P}} \mathbb{E}\left[\mathcal{E}(\widehat{r}_n)\right] \sim \min_{r_n} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\mathcal{E}(r_n)\right] \text{ as } n \to \infty,$$

où l'infimum est pris sur toutes les règles de ranking possibles dépendant de (X_1,Y_1), $\ldots$, (X_n,Y_n). Pour mener une telle étude, on utilise principalement trois hypothèses. On introduit une hypothèse de régularité pour la fonction de régression $\eta : \mathcal{X} \subset \mathbb{R}^d \to [0,1]$ ainsi qu'une condition de régularité sur la loi marginale $\mu(dx)$ et une hypothèse de dispersion de la loi de $\eta(X)$ appelé hypothèse de marge.

L'hypothèse de complexité consiste à considérer des fonctions de régression appartenant à un espace de Hölder de paramètres $(\beta, L, \mathbb{R}^d)$. On considère également deux hypothèses sur la loi marginale: pour l'hypothèse faible la marginale est majorée alors que dans le cas de l'hypothèse forte la marginale est majorée et minorée par des nombres strictement plus grand que zéro. L'hypothèse de marge prend la forme suivante.
Soit $\alpha \in [0, 1]$. On a: $\forall (t, x) \in [0, 1] \times \mathcal{X}$,

$$\mathbb{P}\left\{|\eta(X) - \eta(x)| \leq t\right\} \leq C \cdot t^{\alpha}, \tag{5}$$

pour une constante $C < \infty$.

L'hypothèse (5) ci-dessus est vide si $\alpha = 0$ et de plus en plus restrictive quand α croit. Elle rappelle la condition de marge de Tsybakov, introduite dans [Tsybakov, 2004], qui se revient à (5) avec $1/2$ à la place de $\eta(x)$. Alors que la condition de Tsybakov est liée au comportement de $\eta(X)$ près de $1/2$, l'hypothèse (5) implique le comportement globale de la loi de $\eta(X)$, comme le montre le résultat suivant. Alors que dans le cas de la condition de Tsybakov α peut être très grand jusqu'à l'infini ce que permet de retrouver la condition de marge de Massart [Massart, 2000], l'hypothèse (5) ne peut être remplie que pour $\alpha \leq 1$.

Agrégation par poids exponentiels

La procédure que l'on propose ici est inspiré par [Lecué, 2006] et utilise des poids exponentiels. On montre que la règle de décision obtenue satisfait une inégalité oracle que l'on utilise dans la partie pour montrer des bornes supérieures minimax. Etant donné $r_1, \ldots, r_M$ des règles de ranking, le but d'une procédure d'agrégation est de d'imiter les performances du meilleur de ceux-ci pour l'excès de risque et sous l'hypothèse de marge. On définit la règle de décision agrégée par poids exponentiels par $\tilde{f}_n = \sum_{m=1}^{M} w_m^{(n)} r_m$ où les poids w_m^n valent $w_m^{(n)} = \frac{\exp(\sum_{i \neq j} -Z_{ij} r_m(X_i, X_j))}{\sum_{k=1}^{M} \exp(\sum_{i \neq j} -Z_{ij} r_m(X_i, X_j))}, \forall m = 1, \ldots M.$
On appelle cette procédure *agrégation par poids exponentiels*(AEW). L'idée de cette procédure est de donner plus de poids aux règles de ranking dont le risque empirique est petit de façon à imiter les performances du minimiseurs du hinge risque de ranking empirique (ERM). On montre que la procédure AEW a un excès de risque similaire à celui de l'ERM à un terme additionnel $(\log M)/n$ près. L'intérêt principal de la méthode AEW est qu'elle ne requiert pas une étape de minimisation et qu'elle est moins sensible au sur-apprentissage car la règle de décision agrégée est un mélange pondéré de plusieurs règles de ranking alors que l'ERM implique une seule règle.

On a montré une inégalité oracle pour l'excès de hinge risque de ranking. On suppose que l'hypothèse (5) est vérifiée. On note $\mathcal{C}$ l'enveloppe convexe d'un ensemble fini de règles de décision $\mathcal{F} = \{f_1, \ldots, f_M\}$ à valeurs dans $[-1, 1]$. Soit $\tilde{f}_n$ l'estimateur agrégé défini ci-dessus. Alors, pour tout entier $M \geq 3, n \geq 1$ et tout

$a > 0$, $\tilde{f}_n$ vérifie l'inégalité

$$\mathbb{E}[A(\tilde{f}_n) - A^*] \leq (1+a) \min_{f \in \mathcal{C}} (A(f) - A^*) + C \left(\frac{\log M}{n} \right)^{\frac{\alpha+1}{\alpha+2}},$$

où $C > 0$ est une constante qui ne dépend que de a. Dans [Lecué, 2006], il est montré que la vitesse $\left(\frac{\log M}{n}\right)^{\frac{\alpha+1}{\alpha+2}}$ est optimale au sens minimax dans le cas de la classification binaire. Pour le moment, on n'a pas un tel résultat d'optimalité pour le ranking, cependant la vitesse dans l'inégalité oracle est la même. Cette inégalité oracle est l'outil principal pour obtenir les vitesses minimax adaptatives en utilisant un estimateur basé sur la procédure AEW.

Vitesses rapides pour le ranking binaire

Dans cette partie, on présente les bornes supérieurs minimax pour le ranking bianire dans deux cas, celui sous l'hypothèse faible pour la densité marginale et celui sous l'hypothèse forte. Les estimateurs de la fonction de régression sont les mêmes que ceux utilisés dans le cas de la classification (voir [Lecué, 2006] et [Audibert & A.Tsybakov, 2007]).

Le cas "fort"

Avec les résultats de la partie précédente, on peut établir les bornes supérieures minimax pour le risque de ranking $\inf_{r_n} \sup_{P \in \Sigma} \mathbb{E}[\mathcal{E}(r_n)]$, sous l'ensemble des hypothèses décrites précédemment. Une borne supérieure pour la vitesse minimax se prouve en exhibant une suite de règles de ranking atteignant cette vitesse. Ici, on considère le même estimateur de la fonction de régression que celui étudié dans [Audibert & A.Tsybakov, 2007] i.e. l'estimateur par polynômes locaux. Pour cet estimateur, le maximum du risque de ranking de la règle de ranking plug-in $\widehat{r}_n(x, x') = 2 \cdot \mathbb{I}\{\widehat{\eta}_{n,h_n}(x') > \widehat{\eta}_{n,h_n}(x)\} - 1$ est en $O\left(n^{-\frac{\beta(1+\alpha)}{d+2\beta}}\right)$. Puisque $\alpha \leq 1$, les vitesses plus rapides que n^{-1} ne peuvent pas être atteintes par la règle plug-in $\widehat{r}_n$ définie dans le théorème ci-dessus, en dépit de l'optimalité de l'estimateur associé $\widehat{\eta}_{n,h_n}$. Cependant, pour tout $\alpha \in]0, 1]$, des vitesses rapides peuvent être obtenues (plus rapides que $n^{-1/2}$), dès que la fonction de régression est suffisamment régulière, c'est à dire $\beta > d/2\alpha$. En utilisant l'inégalité oracle précédente, on peut construire un estimateur qui s'adapte aux paramètres de régularité et de marge.

Le cas "faible"

Un résultat important de [Kolmogorov & Tikhomirov, 1961], sur la complexité des classes de Hölder, affirme que le nombre de boules de taille ε pour recouvrir une classe de Hölder de paramètre β est majorée par $\exp(\varepsilon^{-\frac{d}{\beta}})$. On se propose d'agréger, grâce à la procédure AEW, les fonctions qui sont les centres de ces boules pour construire

notre estimateur $\tilde{r}_n^{(\varepsilon,\beta)}$. On montre que l'excès de risque de ranking maximum de la règle agrégée définie par la procédure AEW est borné par $\left(n^{-\frac{\beta(1+\alpha)}{d+\beta(2+\alpha)}}\right)$.

Pour obtenir un estimateur adaptatif à la régularité β et au paramètre de marge α, on agrège des règles de ranking $\tilde{r}_n^{(\varepsilon,\beta)}$ pour (ε, β) sur une grille. On montre que cet estimateur est bien adaptatif aux paramètres et qu'il conserve la même vitesse de convergence. Le nombre de fonctions pour recouvrir une classe de Hölder de paramètre β est exponentiel donc les estimateurs $\tilde{r}^{\varepsilon_n,\beta}$, pour un couple (ε_n, β) donné, ne sont pas facilement implémentable. Cependant la procédure est intéressante d'un point de vue théorique puisqu'elle permet de s'adapter aux paramètres et elle atteint des vitesses rapides dès lors que $\alpha\beta > d$.

Borne inférieure

On a également montré une borne inférieure pour l'excès de risque de ranking dans le cas fort qui a la même vitesse que celle obtenue pour l'estimateur plug-in, sous la restriction que $\alpha\beta \leq 1$. Ce résultat montre que la règle plug-in utilisant les polynômes locaux est optimale dans le cas $\alpha\beta \leq 1$, les vitesses du théorème 7.4.3 étant minimax (et rapide, quand $\alpha\beta > d/2$). Cette preuve repose sur la construction de fonction de régression qui vérifie toutes les hypothèses et sur des idées venant du lemme d'Assouad.

Introduction

The overall purpose of this PhD thesis is to develop and study algorithms which goal is to learn to order observations with unknown labels from a collection of labeled instances.
Suppose that we get a collection of observations, characterized by a set of variables and a label (or a class), belonging to an ordered discrete set. A usual goal is to infer a function that predicts the label of a new observation. This supervised learning task is called ordinal regression. However, there exist many applications where the goal is rather to define an order over the set of observations than a classification. This problem is called the multipartite ranking (or K-partite ranking) task.
In many situations, a natural ordering can be considered over a set of observations. When observations are documents in information retrieval applications, the ordering reflects degree of relevance for a specific query. In order to predict future ordering on new data, the learning process uses past data for which some relevance feedback is provided, such as ratings, say from 0 to 4, from the non relevant to the the extremely relevant. A similar situation occurs in medical applications where decision-making support tools provide a scoring of the population of patients based on diagnostic test statistics in order to rank the individuals according to the advance states of a disease which are described as discrete grades, see [Pepe, 2003], [Dreiseitl *et al.*, 2000], [Edwards *et al.*, 2005], [Mossman, 1999] or [Nakas & Yiannoutsos, 2004] for instance.
While this problem is omnipresent in many applications, theoretical questions related to the learning of prediction functions are still widely open. Several procedures were developed to learn prediction functions for multipartite ranking, but the consistency of the algorithms was not tackled. The main motivation of this manuscript is the understanding of the probabilistic nature of the multipartite ranking problem in order to infer consistent algorithms. Thus, we propose two procedures to achieve this goal, the first one relying on aggregation of prediction functions and the second one based on an recursive approximation scheme.
In the presence of ordinal feedback (*i.e.* ordinal label Y taking a finite number of values $1, \ldots, K$, $K \geq 3$ say), the multipartite ranking task consists in learning how to order temporarily unlabeled observations so as to reproduce as accurately as possible the ordering induced by the labels not observed yet. The natural way to consider this problem is to infer a scoring function s on the learning set that gives a real value for each observation and uses the natural order of the real line to rank the observations. Ideally, as the scoring function s increases, with large probability, we would like to observe, as a majority, the instances with label $Y = 1$ first, those with label $Y = 2$ next, ... and the instances label $Y = K$ getting the higher values.
The issue of ranking data with binary labels, generally termed *bipartite ranking problem*, has recently been the subject of a good deal of attention in the statistical

and machine-learning literature. It leads to the design of novel efficient algorithms fully tailored for the bipartite ranking task (see [Clémençon & Vayatis, 2009b], [Freund *et al.*, 2003] and [Clémençon & Vayatis, 2010] in particular) and gives rise to significant theoretical developments dedicated to this global learning problem (refer to [Agarwal *et al.*, 2005] or [Clémençon *et al.*, 2008] for instance). The extension of related concepts and results to the multipartite context is far from immediate and poses many questions of theoretical or practical nature, not answered yet, see [Flach, 2004] and the references therein. While, in the bipartite framework, the ROC curve (as well as transforms or summaries of the latter such as the celebrated AUC criterion) has provided the "definitive tool" for evaluating the relevance of ranking rules to a certain extent since its introduction in the 40's (*cf* [Green & Swets, 1966]), it is only recently that this functional measure of accuracy has been generalized to the ordinal multi-class setup, leading to a specific notion of "ROC graph" tailored for K-partite ranking, see [Scurfield, 1996]. Until now, the approach to ranking followed by most authors has consisted in optimizing a specific scalar criterion over a (nonparametric) set of ranking/scoring rules and applying the *empirical risk minimization* (ERM) paradigm. The "ranking risk" generally counts the number of "concordant pairs of observations" (*i.e.* the number of pairs of observations that are sorted in the same order as their labels) and takes the form of a U-statistic of degree two, see [Clémençon *et al.*, 2008], [Rudin *et al.*, 2005]. Alternately, in the bipartite framework, it may be a specific functional of the ranks induced by the ranking rule candidate, as in [Rudin, 2006], [Clémençon & Vayatis, 2007] or [Clémençon & Vayatis, 2008].

In the setup, various methods have been proposed in order to develop efficient algorithms. We can cluster these methods in two approaches. In a first group, we find classical statistics methods, which consist in estimating the posterior distributions $\eta_k(x) = \mathbb{P}\{Y = k|X = x\}$ and using these estimators to order the space of observations. This is the case of the Linear Discriminant Analysis (LDA, [Fisher, 1936]) and logistic regression (see [Hastie & Tibshirani, 1990], [Friedman *et al.*, 1998], [Hastie *et al.*, 2001]) that are based on a maximization of the likelihood in a logit type model. An other example is the kernel logistic regression (KLR see for instance [Zhu & Hastie, 2001]) that estimates the posterior distribution by solving a convex optimization problem.
The other approach is based on the optimization of a criterion evaluating the empirical performance (or the risk) of a scoring function. Among those methods, we can cite RankBoost proposed by [Freund *et al.*, 2003] and AdaRank proposed by [Xu & Li, 2007] that are based on the boosting principle (see [Freund & Schapire, 1999]), the methods RankSVM proposed by [Joachims, 2002] and RankRLS proposed by [Pahikkala *et al.*, 2007] that are based on SVM but for different cost functions specifically the hinge loss and the square loss. All these methods are based on the same idea : the optimization of a criterion based on the comparison of pairs of observations. Thus, depending on the method, the

performance criterion can be written as a weighted sum of AUC. More recently, [Waegeman *et al.*, 2008a] proposed an SVM type algorithm to optimize a criterion based on a loss for K-tuple of observations that can be written in function of the discordant K-tuple i.e. the number of K-tuples of observations such that the order induced by the scoring function is not coherent with the observed labels.
Obviously, these two approaches present benefits and drawbacks. While it seems natural to estimate the posterior distributions that describe the model, this approach have some limits. First, the representations of the posterior distributions η_k uses *logit-type* models that may not be adapted to the observed data. Moreover, those methods are impacted by the curse of dimensionality when the space of observations has a high dimension. In this last case, the methods based on the optimization of an empirical criterion permit to obtain more accurate scoring functions. However, those methods provide functions that are good for the global problem but can not ensure to find the best observations of a sample. Nevertheless, in many application, such as information retrieval for instance, we only care about the top of the ranking. Thus, other criteria have been introduced, such as the Discounted Cumulative Gain (DCG see [Cossock & Zhang, 2008], the Average Precision (AP see [Voorhees & Harman, 2005]) and the Expected Reciprocal Rank (ERR) ([Chapelle & Chang, 2011]).

In this manuscript, we characterize the optimal solutions of the ranking problem and we present a functional tool (the ROC surface) that permits to recover the optimal scoring functions. Contrary to the bipartite ranking task, the multipartite ranking task is not a well-posed problem without additional assumption. We use classical theory of stochastic monotonicity ([Lehmann & Romano, 2005]) to make it well-posed. We propose several algorithms based on the maximization of ROC surface that have the assets of both previous approaches. Those algorithms produce scoring functions that mimic the order of the regression function η without its direct estimation in high dimension. The methods introduced in this manuscript rely on generalization of bipartite ranking procedure (see [Clémençon & Vayatis, 2010], [Clémençon *et al.*, 2013a]) to the multipartite ranking setup. All our methods produce piecewise constant scoring functions and can be represented as a binary oriented tree, called ranking tree. The leaves represent cells of the partition of the input space $\mathcal{X}$ and can be visualized.

Organization of the manuscript

Chapter 1 is devoted to the study of existence and characterization of optimal scoring functions for multipartite ranking and performance criteria. In the case of binary ranking this question does not arise, because there is always an optimal scoring function that is the regression function (and its increasing transformations). Given this property, we define optimal functions for multipartite ranking, any function which belongs to the intersection of the sets of optimal functions for sub-binary problems.

We show that this intersection is nonempty if and only if a certain condition on the conditional distributions of classes is respected. This condition is called monotony of the likelihood ratios and is strongly related to another condition existing in the literature (see [Waegeman & Baets, 2011]). Moreover, if this condition is satisfied, it is shown that the optimal functions are scoring functions that order in the same order as the regression function. Then we introduce the ROC surface which is the natural extension of the ROC curve in the ordinal case. We show that the functions that dominate all others in terms of ROC surface, are exactly those which belong to the set of optimal functions for ranking (under monotony of the likelihood ratios). Thus, we obtain a functional criterion to recover the optimal scoring function. This criterion is not easy to handle, we introduce the volume under the ROC surface (VUS) that is real-valued and that criterion is used to compare the numerical performance of ranking algorithms.
In Chapter 2, the scoring function s is fixed and the goal is to estimate the ROC surface and to build confidence regions for the estimators. The distribution of the observations is unknown but we have access to a data sample $\mathcal{D}_n = \{(X_i, Y_i) | i = 1, \ldots, n\}$. We study the difference between the true ROC surface and the estimation based on the sample $\mathcal{D}_n$. A procedure is established for building a non-parametric estimate of the ROC surface and we give a strong approximation of the estimator. Then, through a smooth bootstrap procedure, confidence regions are found for this estimator. Finally, we show examples of confidence regions in a toy example and we compare two scoring functions with real data.
In Chapter 3, the goal is to extend the *Empirical Risk Minimization* (ERM) paradigm, from a practical perspective, to the situation where a natural estimate of the risk is of the form of a K-sample U-statistics, as it is the case in the multipartite ranking problem. Indeed, the numerical computation of the empirical risk is hardly feasible if not infeasible, even for moderate samples sizes. Precisely, it involves averaging $O(n^{d_1+\ldots+d_K})$ terms, when considering a U-statistic of degrees $(d_1, \ldots, d_K)$ based on samples of sizes proportional to n. We propose here to consider a drastically simpler Monte-Carlo version of the empirical risk based on $O(n)$ terms solely, which can be viewed as an *incomplete generalized U-statistic*, and prove that, remarkably, the approximation stage does not damage the ERM procedure and yields a learning rate of order $O_{\mathbb{P}}(1/\sqrt{n})$. Beyond a theoretical analysis guaranteeing the validity of this approach, numerical experiments are displayed for illustrative purpose.

Chapters 4, 5 and 6 present algorithms for solving the multipartite ranking task. In Chapter 4, we introduce the aggregation of binary scoring functions i.e. built from a sample containing only two labels. The procedure is implemented in two steps. The first one is to decompose the problem into binary sub-problems and learn a bipartite scoring function for each sub-problem. The second is to aggregate the obtained functions for each of the sub-problems that make a consensus between them. For this second step, we present the concept of median scoring function based on the Kendall-τ metric. We show that this procedure provides consistent scoring

functions for the VUS criterion if the scoring functions for each binary sub-problems are themselves consistent for the binary sub-problems.
In Chapter 5, we present a new algorithm fully tailored for the multipartite ranking called TreeRank Tournament. This tree-based procedure provides piecewise constant scoring functions. The method consists in recursively split the space of observations such as the ROC surface of the piecewise constant scoring function converge to the optimal ROC surface. Based on this approximation scheme, the TreeRank Tournament is derived by plug-in the empirical version of the unknown quantities.
The purpose of Chapter 6 is the empirical comparison of the proposed algorithm to the state-of-the-art competitors. To this end, we use simulated and real datasets and we recall the empirical criteria to evaluate the scorings functions. The first goal is to compare the set-up of the proposed algorithms in Chapter 4 and 5 and we retain 3 set-up in order to compare to the competitors. The comparison with the competitors (RankBoost ([Rudin *et al.*, 2005]), SVMrank ([Herbrich *et al.*, 2000]) and RLScore [Pahikkala *et al.*, 2007]) highlights the situations where our procedure are well adapted.

In Chapter 7, we return to the bipartite ranking problem and we study the minimax rates. In the standard binary classification setup, under suitable margin assumptions and complexity conditions on the regression function, fast or even super-fast rates can be achieved by plug-in classifiers. In the context of bipartite ranking, no results of this nature has been proved yet. It is the main purpose of the chapter to investigate this issue, by considering bipartite ranking as a nested continuous collection of cost-sensitive classification problems. A global low noise condition is exhibited under which certain ranking rules are proved to achieve fast (but not super-fast) rates over a wide non-parametric class of models.

List of publications

This work has been the subject of the following scientific publications listed below :

- S. Clémençon and S. Robbiano, *Minimax Learning Rates for Bipartite Ranking and Plug-in Rules*, ICML 2011.
- S. Clémençon, S. Robbiano and N. Vayatis, *Ranking Data with Ordinal label: Optimality and Pairwise Aggregation*, Machine Learning, 2013.
- S. Clémençon, S. Robbiano and J. Tressou *Maximal Deviations of Incomplete U-statistics with Applications to Empirical Risk Sampling*, SIAM data Mining, 2013.
- S. Robbiano *Upper Bounds and Aggregation in Bipartite Ranking*, EJS, Accepted.

Other publications are in preparation :

- S. Robbiano, *Consistent Aggregation of Bipartite Rules for Ranking ordinal data*, submitted.
- S. Clémençon and S. Robbiano, *Confidence regions for the ROC Surface via Smoothed Bootstrap*, submitted.
- S. Clémençon and S. Robbiano, *TreeRank Tournament*, submitted.

Part I

Performance measures for multipartite ranking

Chapter 1

Optimality and Performance measures in multipartite ranking

The goal of this chapter is to characterize the optimal elements for the multipartite ranking problem and find a way to recover them. Firstly, we want to define the optimal scoring functions. In bipartite ranking, optimal scoring functions always exist, see [Clémençon & Vayatis, 2009b]. A first intuition suggests that functions which are optimal for all bipartite ranking subproblems simultaneously should be optimal for the global problem. However the intersection of sets of optimal scoring functions for each sub-problem may be empty. We state sufficient conditions for optimality which are called *monotonicity likelihood ratios* conditions. These conditions are similar to the condition introduced in [Waegeman & Baets, 2011] and say that the likelihood ratios have the same monotony. When the likelihood ratios do not satisfy this hypothesis, the set of scoring functions that are optimal for each sub problem is empty. However, it is possible to build collection of distributions such as the *monotonicity likelihood ratios* conditions hold.
In order to solve the multipartite ranking problem, we want to recover the optimal scoring functions. In bipartite ranking, the optimal scoring functions are the ones who maximize receiver operating characteristic (ROC) curve. A natural extension of this tool for multipartite ranking is the so-called ROC surface. The ROC surface was first defined in [Scurfield, 1996] as well as the Volume under the ROC surface (VUS) in the 3-classes case. For the multipartite case, the definition was introduced in [Li & Zhou, 2009]. The ROC surface gives a partial order on the set of scoring functions. We show that the set of optimal scoring functions are exactly the same as the set of scoring functions that maximize the ROC surface in each point so it can be used as a functional criterion for the ranking problem. Its main summarized criterion is the volume under the ROC surface (VUS), we present it version in the general case and we show that the maximizers of this criterion are the optimal functions for the multipartite ranking task.

The rest of the chapter is structured as follows. In Section 1.1, the probabilistic setting is introduced and optimal scoring functions for multipartite ranking are successively defined and characterized. A specific *monotonicity likelihood ratio* condition is stated, which is shown to guarantee the existence of a natural optimal ordering over the input space. The performance metrics, such as the VUS, are the subject matter of Section 1.2. Mathematical proofs are postponed to section 1.4.

1.1 Optimal elements in multipartite ranking

This section is dedicated to the understanding and the characterization of the optimal elements for the multipartite ranking problem. We introduce the notations that are used all along the manuscript and we give several examples to highlight the case when it is possible to have optimal elements.

1.1.1 Probabilistic setup and notations

We consider a black-box system with random input/output pair (X, Y). We assume that the input random vector X takes values over $\mathbb{R}^d$ and the output Y over the ordered discrete set $\mathcal{Y} = \{1, \ldots, K\}$. Here it is assumed that the ordered values of the output Y can be reflected by an ordering over $\mathbb{R}^d$. The case where $K = 2$ is known as the bipartite ranking setup. In the first six chapters of this manuscript, we focus on the case where $K > 2$. Though the objective pursued here is different, the probabilistic setup is exactly the same as that of *ordinal regression*, see subsection 1.1.5 for a discussion of the connections between these two problems. We denote by F_k the conditional distribution function of X given $Y = k$, ϕ_k the density function of the class-conditional distributions of X given $Y = k$ and $\mathcal{X}_k$ its support for $k = 1, \ldots, K$. We also set $p_k = \mathbb{P}\{Y = k\}$, $k \in \{1, \ldots, K\}$, the mixture parameter for class $Y = k$, and $\eta_k(x) = \mathbb{P}\{Y = k \mid X = x\}$ the posterior probability. Set μ the X's marginal distribution and $\phi(x) = p_1\phi_1(x) + \ldots p_K\phi_K(x)$ its density and recall that:

$$\forall k \in \{1, \ldots, K\} \ , \quad \forall x \in \bigcup_{l=1}^{K} \mathcal{X}_l \ , \quad \eta_k(x) = p_k \cdot \frac{\phi_k}{\phi}(x) \quad .$$

The regression function $\eta(x) = \mathbb{E}[Y \mid X = x]$ can be expressed in the following way:

$$\forall x \in \bigcup_{l=1}^{K} \mathcal{X}_l \ , \quad \eta(x) = \sum_{k=1}^{K} k \cdot \eta_k(x) \quad ,$$

as the expectation of a discrete random variable. We shall make use of the following notation for the likelihood ratio of the class-conditional distribution:

$$\forall k, l \in \{1, \ldots, K\} \ , \quad l < k \ , \quad \forall x \in \mathcal{X}_l, \quad \Phi_{k,l}(x) = \frac{\phi_k}{\phi_l}(x) = \frac{p_k}{p_l} \cdot \frac{\eta_k}{\eta_l}(x) \ .$$

Throughout the chapter, we shall use the convention that $u/0 = \infty$ for any $u \in]0, \infty[$ and $0/0 = 0$. $\mathbb{I}\{E\}$ denotes the indicator function of any event E.

1.1.2 Optimal scoring functions

The problem considered in this thesis is to infer an order relationship over $\mathbb{R}^d$ after observing vector data with ordinal labels. For this purpose, we consider real-valued decision rules of the form $s : \mathbb{R}^d \to \mathbb{R}$ called *scoring functions*. In the case of ordinal labels, the main idea is that good scoring functions s are those which assign a high

score $s(X)$ to the observations X with large values of the label Y. We now introduce the concept of optimal scoring function for ranking data with ordinal labels.

Definition 1.1.1. (Optimal scoring function) *An optimal scoring function s^* is a real-valued function such that:*

$$\forall k, l \in \{1, \ldots, K\} \ , \quad l < k \ , \quad \forall x, x' \in \mathcal{X} \ , \quad \Phi_{k,l}(x) < \Phi_{k,l}(x') \Rightarrow s^*(x) < s^*(x') \ .$$

The rationale behind this definition can be understood by considering the case $K = 2$. The class $Y = 2$ should receive higher scores than the class $Y = 1$. In this case, an optimal scoring function s^* should score observations x in the same order as the posterior probability η_2 of the class $Y = 2$ (or equivalently as the ratio $\eta_2/(1 - \eta_2)$). Since $\eta_1(x) + \eta_2(x) = 1$, for all x, it is easy to see that this is equivalent to the condition described in the previous definition (see [Clémençon & Vayatis, 2009b] for details). In the general case ($K > 2$), optimality of a scoring function s^* means that s^* is optimal for all bipartite subproblems with classes $Y = k$ and $Y = l$, with $l < k$.

An important remark is that, in the probabilistic setup introduced above, an optimal scoring function may not exist as shown in the next example.

Example 1. Consider a discrete input space $\mathcal{X} = \{x_1, \ x_2, \ x_3\}$ and $K = 3$. We assume the following joint probability distribution $\mathbb{P}(X = x_i, Y = j) = \omega_{i,j}$ for the random pair (X, Y):

$$\begin{aligned} \omega_{1,1} = \omega_{2,2} = \omega_{3,3} &= 1/2, \\ \omega_{1,2} = \omega_{2,3} = \omega_{3,1} &= 1/3 \\ \omega_{1,3} = \omega_{2,1} = \omega_{3,2} &= 1/6 \ . \end{aligned}$$

Note that in the case of a discrete distribution for X, the density function coincides with mass function and we have $\phi(x) = \mathbb{P}(X = x)$. It is then easy to check that, in this case, there is no optimal scoring function for this distribution in the sense of Definition 1.1.1.

1.1.3 Existence and characterization of optimal scoring functions

The previous example shows that the existence of optimal scoring functions cannot be guaranteed under any joint distribution. Our first important result is the characterization of those distributions for which the family of optimal scoring functions is not an empty set. The next proposition offers a necessary and sufficient condition on the distribution which ensures the existence of optimal scoring functions.

Assumption 1. *For any $k, l \in \{1, \ \ldots, \ K\}$ such that $l < k$, for all $x, x' \in \mathcal{X}$, we have:*

$$\Phi_{k+1,k}(x) < \Phi_{k+1,k}(x') \Rightarrow \Phi_{l+1,l}(x) \leq \Phi_{l+1,l}(x') \ .$$

Proposition 1.1.1. *The following statements are equivalent:*

(1) Assumption 1 holds.

(2) There exists an optimal scoring function s^*.

(3) The regression function $\eta(x) = \mathbb{E}[Y \mid X = x]$ is an optimal scoring function.

(4) For any $k \in \{1, \ldots, K-1\}$, for all $x, x' \in \mathcal{X}_k$, we have:

$$\Phi_{k+1,k}(x) < \Phi_{k+1,k}(x') \Rightarrow s^*(x) < s^*(x') \ .$$

(5) For any $k, l \in \{1, \ldots, K\}$ such that $l < k$, the ratio $\Phi_{k,l}(x)$ is a nondecreasing function of $s^*(x)$.

Assumption 1 characterizes the class distributions for the random pair (X, Y) for which the very concept of an optimal scoring function makes sense. The proposition says that if this condition is not satisfied then the ordinal nature of the labels, when seen through the observation X, is violated. We point out that a related condition, called *ERA ranking representability*, has been introduced in [Waegeman & Baets, 2011], see Definition 2.1 therein. Precisely, it can be easily checked that the condition in the previous proposition means that the collection of (bipartite) ranking rules $\{\Phi_{k+1,k} : \ 1 \leq k < K\}$ is an ERA ranking representable set of ranking rules. Statement (3) suggests that plug-in rules based on the statistical estimation of the regression function η and multiple thresholding of the estimate will offer candidates for practical resolution of multipartite ranking. Such strategies are indeed reminiscent of ordinal logistic regression methods and will be discussed in Part 1.2.7. Statement (4) offers an alternative characterization to Definition 1.1.1 for optimal scoring functions. Statement (5) means that the family of densities of the class-conditional distributions ϕ_k has a *monotone likelihood ratio* (we refer to standard textbooks of mathematical statistics which use this terminology, e.g. [Lehmann & Romano, 2005]).

Proposition 1.1.2. *Under Assumption 1 we necessarily have:*

$$\mathcal{X}_{k'} \cap \mathcal{X}_{l'} \subseteq \mathcal{X}_k \cap \mathcal{X}_l \text{ for any } k, k', l, l' \text{ such that } 1 \leq k' \leq k < l \leq l' \leq K \ .$$

For a better understanding of this proposition, we give a toy example where the distributions are discrete.

Example 2. Consider a discrete input space $\mathcal{X} = \{x_1, \ x_2, \ x_3, \ x_4, \ x_5\}$ and $K = 3$. We assume the following joint probability distribution $\mathbb{P}(X = x_i, Y = j) = \omega_{i,j}$ for the random pair (X, Y):

$$\begin{array}{l} \omega_{1,1} = 1/2, \omega_{1,2} = 1/3, \omega_{1,3} = 1/6, \omega_{1,4} = 2/3, \omega_{1,5} = 1/3 \\ \omega_{1,2} = \omega_{2,2} = \omega_{1,3} = 1/3 \\ \omega_{1,3} = 1/6, \omega_{2,3} = 1/3, \omega_{3,3} = 1/2, \omega_{4,3} = 1/3, \omega_{5,3} = 2/3 \ . \end{array}$$

Here, the support of ϕ_2 is $\mathcal{X}_2 = \{x_1, \ x_2, \ x_3\}$ whereas the support of ϕ_1 and ϕ_3 is $\mathcal{X}$. Here, it is easy to see that the optimal scoring function for "1" versus "2"

rank $x_4 = x_5 < x_1 < x_2 < x_3$, the optimal scoring function for "2" versus "3" rank $x_1 < x_2 < x_3 < x_4 = x_5$ and the optimal scoring function for "1" versus "3" rank $x_1 < x_4 < x_2 < x_5 < x_3$. Clearly in this example we see that the condition on the support is required in order to have a ranking problem well defined.

1.1.4 Examples and counterexamples

It is easy to see that, in absence of Assumption 1, the notion of K-partite ranking hardly makes sense. However, it is a challenging statistical task to assess whether data arise from a mixture of distributions ϕ_k with monotone likelihood ratio. We now provide examples and counterexamples of such cases.

Disjoint supports. Consider the separable case where: $\forall k, l,\ \mathcal{X}_k \cap \mathcal{X}_l = \emptyset$. Then Assumption 1 is clearly fulfilled as for $k \neq l$, we have either $\Phi_{k,l} = 0$ or ∞. It is worth mentioning that in this case, the nature of the K-partite ranking problem does not differ from the multiclass classification setup where there is no order relation between classes.

Exponential families. We recall that $\phi = \sum_{k=1}^{K} p_k \phi_k$ is the marginal distribution function of X. We introduce the following choice for the class-conditional distributions ϕ_k:

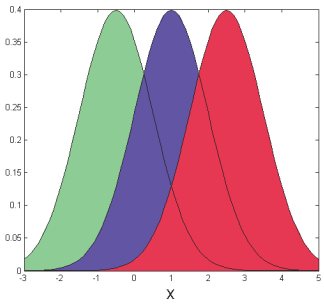
$$\phi_k(x) = \exp\{\kappa(k)T(x) - \psi(k)\}\ f(x)\ , \quad \forall x \in \mathbb{R}^d\ ,$$

where:

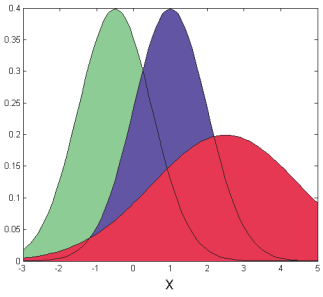
- $\kappa : \{1,\ \ldots,\ K\} \to \mathbb{R}$ is strictly increasing,
- $T : \mathbb{R}^d \to \mathbb{R}$ such that $\psi(k) = \int_{x \in \mathbb{R}^d} \exp\{\kappa(k)T(x)\} f(x) dx < +\infty$, for $1 \leq k \leq K$.

It is easy to check that he family of density functions ϕ_k has the property of monotone likelihood ratio.

1-*D Gaussian distributions.* We consider here a toy example with $K = 3$ and the ϕ_k are univariate Gaussian distributions $\mathcal{N}(m_k,\ \sigma_k^2)$, where m_k is the expectation and σ_k^2 is the variance. Depending on the values of the parameters m_k, σ_k^2, the collection $\{\phi_1, \phi_2, \phi_3\}$ may or may not satisfy the property of having a monotone likelihood ratio. Assume first that the variances are equal, then the property of monotone likelihood ratio is satisfied if and only if either $m_1 \leq m_2 \leq m_3$ or $m_3 \leq m_2 \leq m_1$ (see Figure 1.1-(a)). Figure 1.1-(b) depicts a situation where $m_1 < m_2 < m_3$ and $\sigma_3^2 > \sigma_2^2 = \sigma_1^2$ for which the random observation X does not permit to recover the preorder induced by the output variable. The monotonicity condition is violated for instance at $(x, x') = (-2, 1)$ and there is no optimal scoring function in this case.



(a)
Gaussian class densities fulfilling Assumption 1 : $m_1 = -0.5$, $m_2 = 1$, $m_3 = 2.5$, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$



(b)
Gaussian class densities not fulfilling Assumption 1 : $m_1 = -0.5$, $m_2 = 1$, $m_3 = 2.5$, $\sigma_1^2 = \sigma_2^2 = 1$, $\sigma_3^2 = 2$

Figure 1.1: Two examples of 1-D conditional Gaussian distributions in the case $K = 3$ - class 1 in green, class 2 in blue and class 3 in red.

Uniform noise. Let $t_0 = -\infty < t_1 < \ldots < t_{K-1} < +\infty$ and $g : \mathbb{R}^d \to \mathbb{R}$ a measurable function. We define the output random variable through:

$$Y = \sum_{k=1}^{K} \mathbb{I}\{g(X) + U > t_{k-1}\} \ ,$$

where U is a uniform random variable on some interval of the real line, independent from X. Then it can be easily seen that the class-conditional distributions form a collection with monotone likelihood ratio, provided that t_1 and t_{K-1} both lie inside the interval defined by the essential infimum and supremum of the random variable $g(X) + U$. In this case, any strictly increasing transform of g is an optimal scoring function.

1.1.5 Connection with ordinal regression

Ordinal regression penalize more and more the error of a classifier candidate C on an example (X, Y) as $|C(X) - Y|$ increases. In general, the chosen loss function is of the form $\psi(c, y) = \Psi(|c - y|)$, $(c, y) \in \{1, \ldots, K\}^2$, where $\Psi : \{0, \ldots, K-1\} \to \mathbb{R}_+$ is some nondecreasing mapping. The most commonly used choice is $\Psi(u) = u$, corresponding to the risk $L(C) = \mathbb{E}[|C(X) - Y|]$, referred to as the *expected ordinal regression error* sometimes, *cf* [Agarwal, 2008]. In [Frank & Hall, 2001], they solve the ordinal regression problem by estimating $K-1$ the functions g_k that estimate the probability that the label is over k and then using that the predicting the label that maximize the difference of $g_{k-1} - g_k$ (with $g_0 = 1$). However, another strategy consists in building a scoring function that optimizes the ROC surface or a summary criterion of it like the volume under the ROC surface (VUS) (see [Waegeman *et al.*, 2008a] for instance) and then thresholding the obtained scoring function. In this case, it is shown that the optimal classifier can be built by thresholding the regression function at specific levels $t_0 = 0 < t_1^* < \ldots < t_{K-1}^* < 1 = t_K$, so it is of the form

$C^*(x) = \sum_{k=1}^{K} k \cdot \mathbb{I}\{t^*_{k-1} \leq \eta(x) < t^*_k\}$ when assuming that $\eta(X) = \mathbb{E}[Y \mid X]$ is a continuous r.v. for simplicity. Based on this observation, a popular approach to ordinal regression lies in estimating first the regression function η by an empirical counterpart $\widehat{\eta}$ (through minimization of an estimate of $R(f) = \mathbb{E}[(Y - f(X))^2]$ over a specific class $\mathcal{F}$ of function candidates f, in general) and choosing next a collection $\mathbf{t}$ of thresholds $t_0 = 0 < t_1 < \ldots < t_{K-1} < 1 = t_K$ in order to minimize a statistical version of $L(C_{\mathbf{t}})$ where $C_{\mathbf{t}}(x) = \sum_{k=1}^{K} k \cdot \mathbb{I}\{t_{k-1} \leq \widehat{\eta}(x) < t_k\}$. Such procedures are sometimes termed *regression-based algorithms*, see [Agarwal, 2008]. One may refer to [Kramer *et al.*, 2001] in the case of regression trees for instance. However, the computation of an estimate of the regression is not needed only the order given by the scoring function is important since the thresholds are chosen a posteriori. So any algorithm of multipartite ranking can be used in order to obtain a scoring function and the thresholds can be chosen as previously.

1.2 Performance measures for multipartite ranking

We now turn to the main concepts for assessing performance in the multipartite ranking problem. These concepts are generalizations of the well-known ROC curve and AUC criterion which are popular performance measures for bipartite ranking. We introduce the notation $F_{s,k}$ for the cumulative distribution function (cdf) over the real line $\mathbb{R}$ of the random variable $s(X)$ given the class label $Y = k$:

$$\forall t \in \mathbb{R}\ , \quad F_{s,k}(t) = \mathbb{P}\{s(x) \leq t \mid Y = k\}\ .$$

1.2.1 Reminder on ROC curves

The ROC curve is a crucial notion for understanding the ROC surface. In fact, each pairwise ROC curves can be deduced from the surface (see proposition 1.2.3). This tool has been introduced in the 40's (*cf* [Green & Swets, 1966]) and it provides the definitive tool for evaluating the relevance of scoring function in the bipartite ranking framework. For clarity, we recall here the following definition.

Definition 1.2.1. (ROC CURVE) *The* ROC *curve of a scoring function s with conditional distributions the pair $(F_{s,1}, F_{s,2})$ is the parametrized curve*

$$t \in \mathbb{R} \mapsto \left(\mathbb{P}\left\{s(X) > t \mid Y = 1\right\}, \mathbb{P}\left\{s(X) > t \mid Y = 2\right\}\right).$$

By convention, possible jumps (corresponding to points where the distributions $F_{s,1}(dt)$ and/or $F_{s,2}(dt)$ are degenerate) are connected by line segments, in order to guarantee the continuity of the curve. Equipped with this convention, the ROC curve may be viewed as the graph of a certain non decreasing *càd-làg*[1] mapping

[1] Recall that, by definition, a *càd-làg* function $h : [0,1] \to \mathbb{R}$ is such that $\lim_{s \to t,\ s<t} h(s) = h(t-) < \infty$ for all $t \in]0,1]$ and $\lim_{s \to t,\ s>t} h(s) = h(t)$ for all $t \in [0,1[$. Its completed graph is obtained by connecting the points $(t, h(t-))$ and $(t, h(t))$, when they are not equal, by a vertical line segment and thus forms a continuous curve.

$\alpha \in [0,1] \mapsto \mathrm{ROC}_{\phi_1,\phi_2}(\mathrm{s},\alpha)$, defined by

$$\mathrm{ROC}_{\phi_1,\phi_2}(\mathrm{s},\alpha) = 1 - \mathrm{F}_{\mathrm{s},2} \circ \mathrm{F}_{\mathrm{s},1}^{-1}(1-\alpha)$$

at points α such that $F_{s,1} \circ F_{s,1}^{-1}(1-\alpha) = 1-\alpha$, denoting by $W^{-1}(u) = \inf\{t \in]-\infty,\ +\infty] :\ W(t) \geq u\}$, $u \in [0,1]$, the generalized inverse of any cdf $W(t)$ on $\mathbb{R} \cup \{+\infty\}$. Observe that in absence of plateau, the curve $\alpha \mapsto \mathrm{ROC}_{\phi_2,\phi_1}(\alpha)$ is the image of $\alpha \mapsto \mathrm{ROC}_{\phi_1,\phi_2}(\alpha)$ by the reflection with the line of Eq. "$\beta = \alpha$" as axis. We refer to Appendix A in [Clémençon & Vayatis, 2009b] for a list of properties of ROC curves (see Proposition 17 therein).

1.2.2 ROC surface

Given a scoring function $s :\ \mathbb{R}^d \to \mathbb{R}$, the ROC surface offers a visual display which reflects how the conditional distributions of $s(X)$ given the class label $Y = k$ are separated between each other as $k = 1, \ldots, K$.

Definition 1.2.2. (ROC SURFACE) *Let $K \geq 2$. The ROC surface of a real-valued scoring function s is defined as the plot of the continuous extension of the parametric surface in the unit cube $[0,1]^K$:*

$$\begin{aligned} \Delta &\to [0,1]^K \\ (t_1,\ \ldots,\ t_{K-1}) &\mapsto (\ F_{s,1}(t_1)\ ,\ F_{s,2}(t_2) - F_{s,2}(t_1),\ \ldots,\ 1 - F_{s,K}(t_{K-1}))\ , \end{aligned}$$

where $\Delta = \{(t_1,\ \ldots,\ t_{K-1}) \in \mathbb{R}^{K-1}\ :\ t_1 < \ldots < t_{K-1}\}$.

By "continuous extension", it is meant that discontinuity points, due to jumps or flat parts in the cdfs $F_{s,k}$, are connected by linear segments (parts of hyperplanes). The same convention is considered in the definition of the ROC curve in the bipartite case given in [Clémençon & Vayatis, 2009b]. The ROC surface provides a visual tool for comparing the ranking performance of two scoring functions: we shall say that a scoring function $s(x)$ provides a better ranking than $s'(x)$ when: $\forall (\alpha,\gamma) \in [0,1]^2$,

$$\mathrm{ROC}(\mathrm{s},\alpha,\gamma) \geq \mathrm{ROC}(\mathrm{s}',\alpha,\gamma).$$

This criterion induces a partial order over the space of all scoring functions $\mathcal{S}$.

In the case $K = 3$, on which we restrict our attention from now for simplicity (all results stated in the sequel can be straightforwardly extended to the general situation), the ROC surface thus corresponds to a continuous manifold of dimension 2 in the unit cube of $\mathbb{R}^3$. We also point out that the ROC surface contains the ROC curves of the pairwise problems (ϕ_1,ϕ_2), (ϕ_2,ϕ_3) and (ϕ_1,ϕ_3) which can be obtained as the intersections of the ROC surface with planes orthogonal to each of the axis of the unit cube.

Proposition 1.2.1. (Change of parameterization) *The ROC surface of the scoring function s can be obtained as the plot of the continuous extension of the parametric surface:*

$$\begin{aligned}[0,1]^2 &\to \mathbb{R}^3 \\ (\alpha,\gamma) &\mapsto (\alpha, \mathrm{ROC}(\mathrm{s},\alpha,\gamma), \gamma)\end{aligned}$$

where

$$\mathrm{ROC}(\mathrm{s},\alpha,\gamma) = \left(F_{s,2} \circ F_{s,3}^{-1}(1-\gamma) - F_{s,2} \circ F_{s,1}^{-1}(\alpha)\right)_+ \tag{1.1}$$

$$= \left(\mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s}, 1-\alpha) - \mathrm{ROC}_{\mathrm{F}_3,\mathrm{F}_2}(\mathrm{s},\gamma)\right)_+ \ , \tag{1.2}$$

with the notation $u_+ = \max(0,\ u)$, for any real number u.

We point out that, in the case where s has no capacity to discriminate between the three distributions, *i.e.* when $F_{s,1} = F_{s,2} = F_{s,3}$, the ROC surface boils down to the surface delimited by the triangle that connects the points $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$, we then have $\mathrm{ROC}(\mathrm{s},\alpha,\gamma) = 1-\alpha-\gamma$. By contrast, in the separable situation (see Section 1.1.4), the optimal ROC surface coincides with the surface of the unit cube $[0,1]^3$.

The next lemma characterizes the support of the function whose plot corresponds to the ROC surface (see Figure 1.2.2).

Lemma 1.2.2. *For any $(\alpha,\gamma) \in [0,1]^2$, the following statements are equivalent:*

1. $\mathrm{ROC}(\mathrm{s},\alpha,\gamma) > 0$
2. $\mathrm{ROC}_{\phi_1,\phi_3}(\mathrm{s}, 1-\alpha) > \gamma$.

We denote $\mathcal{I}_s$ the set where the ROC *surface is non-negative i.e.*

$$\mathcal{I}_s \stackrel{def}{=} \{(\alpha,\gamma) \in [0,1]^2 :\ \gamma \leq \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s}, 1-\alpha)\}.$$

Other notions of ROC surface have been considered in the literature, depending on the learning problem considered and the goal pursued. In the context of multi-class pattern recognition, they provide a visual display of classification accuracy, as in [Ferri *et al.*, 2003] (see also [Fieldsend & Everson, 2005], [Fieldsend & Everson, 2006] and [Hand & Till, 2001]) from a *one-versus-one* angle or in [Flach, 2004] when adopting the *one-versus-all* approach. The concept of ROC analysis described above is more adapted to the situation where a natural order on the set of labels exists, just like in ordinal regression, see [Waegeman *et al.*, 2008c].

Notice that $F_{s,3}(t_3) = 1$ and $F_{s,1}(t_0) = 0$ and that the coordinates of the point (1.2.2) coincides with the diagonal entries of the *confusion matrix* of the classification rule defined by thresholding $s(X)$ at the levels t_k, $1 \leq k < 3$,

$$C_{\mathbf{t}}(s) = \sum_{k=1}^{3} k \cdot \mathbb{I}\{t_{k-1} < s(X) \leq t_k\}.$$

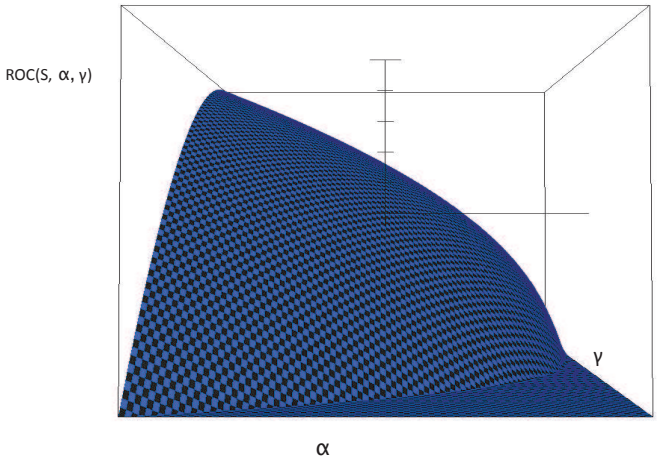


Figure 1.2: Plot of the ROC surface of a scoring function.

We have indeed $\mathbb{P}\{C_{\mathbf{t}}(s(X)) = k \mid Y = k\} = F_{s,k}(t_k) - F_{s,k}(t_{k-1})$ for all k in $\{1, 2, 3\}$.

Remark 1.2.1. (On graph conventions.) *It should be pointed up that, in the case $K = 2$, the* ROC *surface defined above does not coincide with the* ROC *curve defined in subsection 1.2.1 (see Definition 1.2.1) but with its image by the transformation $(\alpha, \beta) \in [0, 1]^2 \mapsto (1 - \alpha, \beta)$.*

The next result summarizes several crucial properties of ROC surfaces.

Proposition 1.2.3. (Properties of the ROC surface)

1. **Intersections with the facets of the** ROC **space.** *The intersection of the* ROC *surface $\{(\alpha, \mathrm{ROC}(\mathrm{s}, \alpha), \gamma)\}$ with the plane of Eq. "$\alpha = 0$" coincides with the curve $\{(\beta, \mathrm{ROC}_{\mathrm{F}_2,\mathrm{F}_3}(\mathrm{s}, \beta))\}$ up to the transform*

   $(\beta, \gamma) \in [0, 1]^2 \mapsto \psi(\beta, \gamma) = (1-\beta, \gamma)$*, that with the plane of Eq. "$\beta = 0$" corresponds to the image of the curve $\{(\alpha, \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_3}(\mathrm{s}, \alpha))\}$ by the mapping $\psi(\alpha, \gamma)$ and that with the plane of Eq. "$\gamma = 0$" to the image of $\{(\alpha, \mathrm{ROC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s}, \alpha))\}$ by the transform $\psi(\alpha, \beta)$.*

2. **Invariance.** *For any strictly increasing function $T : \mathbb{R} \cup \{+\infty\} \to \mathbb{R} \cup \{+\infty\}$, we have, for all $(\alpha, \gamma) \in [0, 1]^2$:*

$$\mathrm{ROC}(\mathrm{T} \circ \mathrm{s}, \alpha, \gamma) = \mathrm{ROC}(\mathrm{s}, \alpha, \gamma).$$

3. **Concavity.** *If the likelihood ratios $\phi_2/\phi_1(u)$ and $\phi_3/\phi_2(u)$ are both (strictly) increasing transformations of a function $T(u)$, then the* ROC* *surface is (strictly) concave.*

4. **Flat parts.** *If the likelihood ratios $\phi_2/\phi_1(u)$ and $\phi_3/\phi_2(u)$ are simultaneously constant on some interval in the range of the scoring function Z, then the* ROC *surface will present a flat part (i.e. will be a part of a plane) on the corresponding domain.*

5. **Differentiability.** *Assume that the distributions $F_1(dx)$, $F_2(dx)$ and $F_3(dx)$ are continuous. Then, the* ROC *surface of a scoring function s is differentiable if and only if the conditional distributions $F_{s,1}(dt)$, $F_{s,2}(dt)$ and $F_{s,3}(dt)$ are continuous. In such a case, denoting by $f_{s,1}$, $f_{s,2}$ and $f_{s,3}$ the corresponding densities, we have in particular: $\forall(\alpha,\gamma) \in \mathcal{I}_s$*

   if $f_{s,1}(F_{s,1}^{-1}(\alpha)) > 0$, $\frac{\partial}{\partial\alpha}\mathrm{ROC}(\mathrm{s},\alpha,\gamma) = -\frac{\mathrm{f_{s,2}}}{\mathrm{f_{s,1}}}\left(\mathrm{F_{s,1}^{-1}}(\alpha)\right)$,

   and if $f_{s,3}(F_{s,3}^{-1}(1-\gamma)) > 0$, $\frac{\partial}{\partial\gamma}\mathrm{ROC}(\mathrm{s},\alpha,\gamma) = -\frac{\mathrm{f_{s,2}}}{\mathrm{f_{s,3}}}\left(\mathrm{F_{s,3}^{-1}}(1-\gamma)\right)$.

Alternative ROC **graph.**

Another way of quantifying the ranking accuracy of a scoring function in the multi-class setting is to evaluate its ability to discriminate between X's conditional distributions given $Y \leq k$ and $Y > k$ respectively, which we denote $h_k(x)$ and $g_k(x)$, for $k \in \{1, \ldots, K-1\}$. This boils down to plot the graph of the mapping $\alpha \in [0,1] \mapsto (\mathrm{ROC}_{\mathrm{h_1,g_1}}(\mathrm{s},\alpha), \ldots, \mathrm{ROC}_{\mathrm{h_{K-1},g_{K-1}}}(\mathrm{s},\alpha))$. It straightforwardly follows from the stipulated monotonicity hypothesis (*cf* Assumption 1) that the curve related to $s^* \in \mathcal{S}^*$ dominates the curve of any other scoring function s in the coordinate-wise sense: $\mathrm{ROC}_{\mathrm{h_k,g_k}}(\mathrm{s},\alpha) \leq \mathrm{ROC}_{\mathrm{h_k,g_k}}(\mathrm{s}^*,\alpha)$ for all $\alpha \in [0,1]$, $1 \leq k < K$. The likelihood ratio $g_k/h_k(X)$ is indeed a non decreasing function of $s^*(X)$, see Theorem 3.4.1 in [Lehmann & Romano, 2005] for instance. However, with such a functional representation of ranking performance, one loses an attractive advantage, the insensitivity to the class probabilities p_k. Indeed, the distributions $h_k(x)$ and $g_k(x)$ depend on the latter, they can be expressed as $\sum_{l\leq k} p_l f_l(x)/(\sum_{l\leq k} p_l)$ and $\sum_{l>k} p_l f_l(x)/(\sum_{l>k} p_l)$ respectively.

On the ROC **surface of a classification rule.**

Let $C : \mathcal{X} \to \mathcal{Y}$ be a classifier. We call $\alpha_{i,j} = \mathbb{P}\{C(X) = i \mid Y = j\}$ for all i,j in $\{1,2,3\}^2$ and $\mathcal{M}(C) = \{\alpha_{i,j}\}$ the confusion matrix of the classifier C. We point out that, with the convention previously introduced, the ROC surface of a classifier $C : \mathcal{X} \to \{1,2,3\}$ is the polyhedron with vertices $(0,0,1)$, $(0,\alpha_{2,1},1-\alpha_{3,1})$, $(0,1-\alpha_{2,3},\alpha_{3,3})$, $(0,1,0)$, $(\alpha_{1,1},0,1-\alpha_{3,1})$, $(\alpha_{1,1},\alpha_{2,2},\alpha_{3,3})$, $(\alpha_{1,1},1-\alpha_{2,1},0)$, $(1-\alpha_{1,3},0,\alpha_{3,3})$, $(1-\alpha_{1,3},\alpha_{2,3},0)$ and $(1,0,0)$, where $\alpha_{k,l} = \mathbb{P}\{C(X) = l \mid Y = k\}$.

We underline that the confusion matrix $\mathcal{M}(C) = \{\alpha_{k,l}\}$ can be fully recovered from this geometric solid, which is actually a decahedron when the matrix $\mathcal{M}(C)$ has no null entry. Observe finally that this graphic representation of $\mathcal{M}(C)$ differs from that which derives from the multi-class notion of ROC analysis proposed in [Ferri *et al.*, 2003]. In the latter case, the ROC space is defined as $[0,1]^6$ and $\mathcal{M}(C)$ is represented by the point with coordinates $(\alpha_{1,2}, \alpha_{1,3}, \alpha_{2,1}, \alpha_{2,3}, \alpha_{3,1}, \alpha_{3,2})$. Notice incidentally that the latter concept of ROC analysis is more general in the sense that it permits to visualize the performance of $K(K-1)/2$ classifiers involved in a *one-versus-one* classification method. As for the ROC curve, each point of the ROC surface is associated to the performances of a classifier. Because the definition of the ROC surface only uses the true class rate (i.e. the diagonal of the confusion matrix), two non-equal classifiers can be represented by the same point in the ROC space of dimension 3. That is why, this ROC surface is sometimes called the simplified ROC surface. In opposition, the full ROC space (dimension 6) uses every false class rates (i.e. the extra diagonal of the confusion matrix).

1.2.3 ROC-optimality and optimal scoring functions

The ROC surface provides a visual tool for assessing ranking performance of a scoring function. The next theorem provides a formal statement to justify this practice.

Theorem 1.2.4. *The following statements are equivalent:*

1. *Assumption 1 is fulfilled and s^* is an optimal scoring function in the sense of Definition 1.1.1.*
2. *We have, for any scoring function s and for all $(\alpha, \gamma) \in [0,1]^2$,*

$$\mathrm{ROC}(\mathrm{s}, \alpha, \gamma) \leq \mathrm{ROC}(\mathrm{s}^*, \alpha, \gamma) \ .$$

A nontrivial byproduct of the proof of the previous theorem is that optimizing the ROC surface amounts to simultaneously optimizing the ROC curves related to the two pairs of distributions (ϕ_1, ϕ_2) and (ϕ_2, ϕ_3).

The theorem indicates that optimality for scoring functions in the sense of Definition 1 is equivalent to optimality in the sense of the ROC surface. Therefore, the ROC surface provides a complete characterization of the ranking performance of a scoring function in the multipartite ranking problem. This theorem is the main justification for introducing the ROC surface.

We now introduce the following notations: for any $\alpha \in [0,1]$ and any scoring function s,

- the quantile of order $(1-\alpha)$ of the conditional distribution of the random variable $s(X)$ given $Y = k$:

$$Q^{(k)}(s, \alpha) = F_{s,k}^{-1}(1-\alpha) \ ,$$

- the level set of the scoring function s with the top elements of class $Y = k$:

$$R_{s,\alpha}^{(k)} = \{x \in \mathcal{X} | s(x) > Q^{(k)}(s, \alpha)\} \ .$$

Proposition 1.2.5. *Suppose that Assumption 1 is fulfilled and consider s^* an optimal scoring function in the sense of Definition 1.1.1. Assume also that $\eta(X)$ is a continuous random variable, then we have: $\forall(\alpha, \gamma) \in [0,1]^2$:*

$$\mathrm{ROC}^*(\mathrm{s}^*, \alpha, \gamma) - \mathrm{ROC}(\mathrm{s}, \alpha, \gamma) \leq \mathbb{I}\{\gamma \leq \mathrm{ROC}_{\phi_1,\phi_3}(\mathrm{s}^*, 1-\alpha)\} \cdot (\Theta_1(\mathrm{s}, \alpha) + \Theta_2(\mathrm{s}, \gamma)) \, ,$$

where

$$\begin{aligned}
\Theta_1(s, \alpha) &= \frac{\mathbb{I}\{\alpha \neq 0\}}{p_2 Q^{(1)}(\eta_1, \alpha)} \mathbb{E}\left[\left|\eta_1(x) - Q^{(1)}(\eta_1, \alpha)\right| \cdot \mathbb{I}\{R_{s^*,\alpha}^{(1)} \Delta R_{s,\alpha}^{(1)}\}\right] , \\
\Theta_2(s, \gamma) &= \frac{\mathbb{I}\{\gamma \neq 1\}}{p_2 Q^{(3)}(\eta_3, 1-\gamma)} \mathbb{E}\left[\left|\eta_3(X) - Q^{(3)}(\eta_3, 1-\gamma)\right| \cdot \mathbb{I}\{R_{s^*,1-\gamma}^{(3)} \Delta R_{s,1-\gamma}^{(3)}\}\right] \ .
\end{aligned}$$

We have used the notation $A \Delta B = (A \setminus B) \cup (B \setminus A)$ for the symmetric difference between sets A and B.

The previous proposition provides a key inequality for the statistical results developed in the sequel. Moreover, it indicates the nature of the ranking problem, indeed if one can recover all the conditional regions of the regression function η, then the problem of multipartite ranking is solved. This task is impossible since the distribution functions are unknown but it inspires the algorithm introduced in chapter 5.

1.2.4 Reminder on the AUC criterion

In the bipartite case, a standard summary of ranking performance is the Area Under an ROC Curve (or AUC). We introduce the definition for the AUC of a scoring function s.

Definition 1.2.3. (Area Under the ROC Curve) *The AUC of a real-valued scoring function s is defined as :*

$$\mathrm{AUC}(\mathrm{s}) = \int_0^1 \mathrm{ROC}(\mathrm{s}, \alpha) \, \mathrm{d}\alpha \ .$$

We recall the probabilistic form of AUC of the bipartite ranking problem with the pair of distributions (ϕ_k, ϕ_{k+1}):

Proposition 1.2.6. *([Clémençon et al., 2011a]) Let X_1 and X_2 independent random variables with distribution ϕ_k and ϕ_{k+1} respectively. We set:*

$$\mathrm{AUC}_{\phi_\mathrm{k},\phi_{\mathrm{k+1}}}(\mathrm{s}) = \mathbb{P}\left\{\mathrm{s}(\mathrm{X}_1) < \mathrm{s}(\mathrm{X}_2)\right\} + \frac{1}{2}\mathbb{P}\left\{\mathrm{s}(\mathrm{X}_1) = \mathrm{s}(\mathrm{X}_2)\right\} \ .$$

This proposition is essential for the empirical evaluation of the ranking performance. Indeed, given a testing sample $\mathcal{D} = \{(X_i, Y_i) | i = 1, \ldots, n\}$ where the random couples (X_i, Y_i) are i.i.d. of independent copies of a random pair (X, Y) where Y is drawn from the distribution $p_k\delta_k + p_{k+1}\delta_{k+1}$, F_k being the conditional distribution of X given $Y = k$ and $p_k = \mathbb{P}\{Y = k\}$, one can compute the empirical AUC using the following statistic :

$$\widehat{\mathrm{AUC}}_{\phi_k,\phi_{k+1}}(s) = \frac{1}{n_k n_{k+1}} \sum_{i=1}^{n_k} \sum_{j=1}^{n_{k+1}} \mathbb{I}\{s(X_i) < s(X_j)\} + \frac{1}{2}\mathbb{I}\{s(X_i) = s(X_j)\}$$

where $n_k = \sum_{i=1}^{n} \mathbb{I}\{Y_i = k\}$ and $n_{k+1} = \sum_{i=1}^{n} \mathbb{I}\{Y_i = k+1\}$. The empirical AUC is then a U-statistic (see [Depecker, 2010] for more details).

We now state the result which establishes the relevance of AUC as an optimality criterion for the bipartite ranking problem.

Proposition 1.2.7. *Fix $k \in \{1, \ldots, K-1\}$ and consider s_k^* a pairwise optimal scoring function according to Definition 4.1.3. Then we have, for any scoring function:*

$$\mathrm{AUC}_{\phi_k,\phi_{k+1}}(s) \leq \mathrm{AUC}_{\phi_k,\phi_{k+1}}(s_k^*) \ .$$

Moreover, we denote the maximal value of the AUC for the bipartite (ϕ_k, ϕ_{k+1}) ranking problem by: $\mathrm{AUC}^*_{\phi_k,\phi_{k+1}} = \mathrm{AUC}_{\phi_k,\phi_{k+1}}(s_k^*)$.

This proposition validates the choice of the AUC as a real criterion for the evaluation of scoring function. The next sections show how to extend these results in the case where $K = 3$.

1.2.5 Volume under the ROC surface (VUS)

In a similar manner, one may consider the *volume under the* ROC *surface* (VUS in abbreviated form) in the three-class framework. We follow here [Scurfield, 1996] but we mention that other notions of ROC surface can be found in the literature, leading to other summary quantities, also referred to as VUS, such as those introduced in [Hand & Till, 2001].

Definition 1.2.4. (Volume Under the ROC Surface) *We define the VUS of a real-valued scoring function s as:*

$$\mathrm{VUS}(s) = \int_0^1 \int_0^1 \mathrm{ROC}(s, \alpha, \gamma) \ d\alpha d\gamma \ .$$

An alternative expression of VUS can be derived with a change of parameters:

$$\mathrm{VUS}(s) = \int_0^1 \mathrm{ROC}_{\phi_1,\phi_2}(s, 1-\alpha)\mathrm{ROC}_{\phi_1,\phi_3}(s, 1-\alpha) \ d\alpha \\ - \int_0^1 \mathrm{ROC}_{\phi_3,\phi_2}(s, \gamma) \left(1 - \mathrm{ROC}_{\phi_3,\phi_1}(s, \gamma)\right) \ d\gamma \ .$$

The next proposition describes two extreme cases.

Proposition 1.2.8. *Consider a real-valued scoring function s.*

1. *If $F_{s,1} = F_{s,2} = F_{s,3}$, then* $\mathrm{VUS(s)} = 1/6$.
2. *If the density functions of $F_{s,1}$, $F_{s,2}$, $F_{s,3}$ have disjoint supports, then* $\mathrm{VUS(s)} = 1$.

Like the AUC criterion, the VUS can be interpreted in a probabilistic manner. We recall the following result i.e. that the VUS can be viewed as the probability of well classifying a random triplet (with random break of the ties).

Proposition 1.2.9. ([Scurfield, 1996]) *For any scoring function $s \in \mathcal{S}$, we have:*

$$\begin{aligned}\mathrm{VUS(s)} = \mathbb{P}\left\{\mathrm{s(X_1)} < \mathrm{s(X_2)} < \mathrm{s(X_3)} | \mathrm{Y_1} = 1, \mathrm{Y_2} = 2, \mathrm{Y_3} = 3\right\} \\ + \frac{1}{2}\mathbb{P}\left\{s(X_1) = s(X_2) < s(X_3) | Y_1 = 1, Y_2 = 2, Y_3 = 3\right\} \\ + \frac{1}{2}\mathbb{P}\left\{s(X_1) < s(X_2) = s(X_3) | Y_1 = 1, Y_2 = 2, Y_3 = 3\right\} \\ + \frac{1}{6}\mathbb{P}\left\{s(X_1) = s(X_2) = s(X_3) | Y_1 = 1, Y_2 = 2, Y_3 = 3\right\},\end{aligned}$$

where (X_1, Y_1), (X_2, Y_2) and (X_3, Y_3) denote independent copies of the random pair (X, Y).

In the case where the distribution of s is continuous, the last three terms on the right hand side vanish and the VUS boils down to the probability that, given three random instances X_1, X_2 and X_3 with respective labels $Y_1 = 1$, $Y_2 = 2$ and $Y_3 = 3$, the scoring function s ranks them in the right order.

As for the AUC, the VUS can be estimated with an U-statistic of order 3 with kernel

$$\begin{aligned}q(X_1, X_2, X_3) = \mathbb{I}\left\{s(X_1) < s(X_2) < s(X_3)\right\} + \frac{1}{2}\mathbb{I}\left\{s(X_1) = s(X_2) < s(X_3)\right\} \\ + \frac{1}{2}\mathbb{I}\left\{s(X_1) < s(X_2) = s(X_3)\right\} + \frac{1}{6}\mathbb{I}\left\{s(X_1) = s(X_2) = s(X_3)\right\}.\end{aligned}$$

Remark 1.2.2. *Let $K \geq 2$ and consider independent random variables $X_1, \ldots, X_K$ defined on the same probability space, taking their values in the same space $\mathcal{X}$ and drawn from distributions $F_1, \ldots, F_K$ fulfilling* **Assumption 1**. *Consider a scoring function $s : \mathcal{X} \to \mathbb{R}$. For any $k \in \{1, \ldots, K-1\}$, set $\mathcal{E}_k(0) = \{s(X_k) < s(X_{k+1})\}$ and $\mathcal{E}_k(1) = \{s(X_k) = s(X_{k+1})\}$.*

Proposition 1.2.10. *The volume of its* ROC *manifold is given by:*

$$\mathrm{VUS(s)} = \sum_{\mathrm{u} \in \{0,\,1\}^{\mathrm{K}-1}} \frac{\mathbb{P}\{\mathcal{E}_1(\mathrm{u_1}) \cap \cdots \cap \mathcal{E}_{\mathrm{K}-1}(\mathrm{u_{K-1}})\}}{\mathcal{D}_{\mathrm{u}}}, \tag{1.3}$$

where $\mathcal{D}_u = (1+(\tau_1-2)\mathbb{I}\{\tau_1 > 1\}) \times \prod_{j=2}^{K_u}(\tau_j - \tau_{j-1} - 1) \times (1+(K-2-\tau_{K_u})\mathbb{I}\{\tau_{K_u} < K-1\})$, $K_u = K - 1 - \sum_{k=1}^{K-1} u_k$, $\tau_1 = \inf\{k \geq 1 : \ u_k = 0\}$ *and* $\tau_j = \inf\{k > \tau_{j-1} : u_k = 0\}$ *for* $1 < j \leq K_u$.

Note that the size of g_u *varies with* u. *For the sake of clarity, let us give an example: if* $u = (1,1,0,0,1,0,1,1,1,0)$ *then* $g_u = (2,1,3)$. $\frac{1}{\prod_{i=1}^{|g_u|_1}(g_u(i)+1)!}$ *is the probability of finding the right order by breaking ties randomly and uniformly. Moreover, there exists a fast algorithm to calculate the empirical version of the VUS* ($O(n\ln(n))$) *[Waegeman et al., 2008b] where* n *is the number of observations, when there is no tie. A slight modification of this algorithm can handle the case with ties (see 1.5).*

1.2.6 VUS-optimality

We now consider the notion of optimality with respect to the VUS criterion and provide expressions of the deficit of VUS for any scoring function which highlights the connection with AUC maximizers for the bipartite subproblems.

Proposition 1.2.11. (VUS OPTIMALITY) *Under Assumption 1, we have, for any real-valued scoring function* s *and any optimal scoring function* s^**:*

$$\mathrm{VUS(s)} \leq \mathrm{VUS(s^*)} \ .$$

We denote the maximal value of the VUS by $\mathrm{VUS^*} = \mathrm{VUS(s^*)}$

This result shows that optimal scoring functions in the sense of Definition 1.1.1 coincide with optimal elements in the sense of VUS. This simple statement grounds the use of empirical VUS maximization strategies for the multipartite ranking problem.

When the Assumption 1 is not fulfilled, the VUS can still be used as a performance criterion, both in the multiclass classification context ([Landgrebe & Duin, 2006], [Ferri *et al.*, 2003]) and in the ordinal regression setup ([Waegeman *et al.*, 2008c]). However, the interpretation of maximizers of VUS as optimal orderings is highly questionable. For instance, in the situation described in Example 1, one may easily check that, when $\omega_{1,1} = 4/11$, $\omega_{1,2} = 6/11$, $\omega_{1,3} = \omega_{3,1} = 1/11$, $\omega_{2,1} = \omega_{2,2} = 3/11$ and $\omega_{2,3} = \omega_{3,2} = \omega_{3,3} = 5/11$, the maximum VUS (equal to 0.2543) is reached by the scoring function corresponding to strict orders $\prec$ and $\prec'$, such that $x_3 \prec x_2 \prec x_1$ and $x_2 \prec' x_3 \prec' x_1$ respectively, both at the same time.

The next result makes clear that if a scoring function s solves simultaneously all the bipartite ranking subproblems then it also solves the global multipartite ranking problem. For simplicity, we present the result in the case $K = 3$.

Theorem 1.2.12. (DEFICIT OF VUS) *Suppose that Assumption 1 is fulfilled. Then, for any scoring function s and any optimal scoring function s^*, we have*

$$\begin{aligned}\mathrm{VUS}(s^*) - \mathrm{VUS}(s) &\leq \left(\mathrm{AUC}^*_{\phi_1,\phi_2} - \mathrm{AUC}_{\phi_1,\phi_2}(s)\right) + \left(\mathrm{AUC}^*_{\phi_2,\phi_3} - \mathrm{AUC}_{\phi_2,\phi_3}(s)\right) \\ &\leq \frac{2}{3}\left(\left(\mathrm{AUC}^*_{\phi_1,\phi_2} - \mathrm{AUC}_{\phi_1,\phi_2}(s)\right) + \left(\mathrm{AUC}^*_{\phi_2,\phi_3} - \mathrm{AUC}_{\phi_2,\phi_3}(s)\right)\right. \\ &\qquad \left. + \left(\mathrm{AUC}^*_{\phi_1,\phi_3} - \mathrm{AUC}_{\phi_1,\phi_3}(s)\right)\right).\end{aligned}$$

This theorem is very important for the chapter 4 where we present an algorithm that aggregates scoring functions that are solutions of the bipartite sub-problems. Thanks to this theorem, we are able to prove VUS-consistency for the procedure.

1.2.7 Approaches to K-partite ranking

In this section, we mention some approaches to multipartite ranking.

Empirical VUS maximization

The first approach extends the popular principle of empirical risk minimization, see [Vapnik, 1999]. For K-partite ranking, this program has been carried out in [Rajaram & Agarwal, 2005] with an accuracy measure based on the loss function $(Y - Y')_+^{\xi}(\mathbb{I}\{s(X) < s(X')\} + (1/2)\cdot\mathbb{I}\{s(X) = s(X')\})$, with $\xi \geq 0$. In our setup, the idea would be to optimize a statistical counterpart of the unknown functional $\mathrm{VUS}(.)$ over a class $\mathcal{S}_1$ of candidate scoring rules. Based on the training dataset $\mathcal{D}_n$, a natural empirical counterpart of $\mathrm{VUS}(s)$ is the three-sample U-statistic

$$\widehat{\mathrm{VUS}}_n(s) = \frac{1}{n_1 n_2 n_3} \sum_{1\leq i,\, j,\, k\leq n} h_s(X_i, X_j, X_k)\cdot\mathbb{I}\{Y_i = 1,\ Y_j = 2,\ Y_k = 3\}, \quad (1.4)$$

with kernel given by

$$\begin{aligned}h_s(x_1,x_2,x_3) = \mathbb{I}\{s(x_1) < s(x_2) < s(x_3)\} + \frac{1}{2}\mathbb{I}\{s(x_1) = s(x_2) < s(x_3)\} + \\ \frac{1}{2}\mathbb{I}\{s(x_1) < s(x_2) = s(x_3)\} + \frac{1}{6}\mathbb{I}\{s(x_1) = s(x_2) = s(x_3)\},\end{aligned}$$

for any $(x_1, x_2, x_3) \in \mathcal{X}^3$. The computational complexity of empirical VUS calculation is investigated in [Waegeman *et al.*, 2008b].

The theoretical analysis shall rely on concentration properties of U-processes in order to control the deviation between the empirical and theoretical versions of the VUS criterion uniformly over the class $\mathcal{S}_1$. Such an analysis was performed in the bipartite case in [Clémençon *et al.*, 2008] and we extends it in the K-partite case in Chapter 3. In contrast, algorithmic aspects of the issue of maximizing the empirical VUS criterion (or a concave surrogate) are much less straightforward and the question of extending optimization strategies such as those introduced in [Clémençon & Vayatis, 2009b] or [Clémençon & Vayatis, 2010] requires, for instance, significant methodological progress.

Plug-in scoring rule

As shown by Proposition 1.1.1, when Assumption 1 is fulfilled, the regression function η is an optimal scoring function. The *plug-in* approach consists of estimating the latter and uses the resulting estimate as a scoring rule. For instance, one may estimate the posterior probabilities $(\eta_1(x), \ \ldots, \ \eta_K(x))$ by an empirical counterpart $(\widehat{\eta}_1(x), \ \ldots, \ \widehat{\eta}_K(x))$ based on the training data and consider the ordering on $\mathbb{R}^d$ induced by the estimator $\widehat{\eta}(x) = \sum_{k=1}^{K} k\widehat{\eta}_k(x)$. We refer to [Clémençon & Vayatis, 2009a] and Chapter 7 of this manuscript for preliminary theoretical results based on this strategy in the bipartite context and [Audibert & A.Tsybakov, 2007] for an account of the plug-in approach in binary classification. It is expected that an accurate estimate of $\eta(x)$ will define a ranking rule similar to the optimal one, with nearly maximal VUS. As an illustration of this approach, the next result relates the *deficit of* VUS of a scoring function $\widehat{\eta}$ to its $L_1(\mu)$-error as an estimate of η. We assume for simplicity that all class-conditional distributions have the same support.

Proposition 1.2.13. *Suppose that Assumption 1 is fulfilled. Let $\widehat{\eta}$ be an approximant of η. Assume that both the random variables $\eta(X)$ and $\widehat{\eta}(X)$ are continuous. We have:*

$$\mathrm{VUS}^* - \mathrm{VUS}(\widehat{\eta}) \leq \frac{\mathrm{p}_1 + \mathrm{p}_3}{\mathrm{p}_1\mathrm{p}_2\mathrm{p}_3} \cdot \mathbb{E}\left[|\eta(\mathrm{X}) - \widehat{\eta}(\mathrm{X})|\right]$$

This result reveals that a $L_1(\mu)$-consistent estimator, *i.e.* an estimator $\widehat{\eta}_n$ such that $\mathbb{E}[|\eta(X) - \widehat{\eta}_n(X)|]$ converges to zero in probability as $n \to \infty$, yields a VUS-consistent ranking procedure. However, from a practical perspective, such procedures should be avoided when dealing with high-dimensional data, since they are obviously confronted to the curse of dimensionality.

1.3 Conclusion

In this chapter, we present theoretical work on ranking data with ordinal labels. In the first part of the chapter, the issue of optimality is tackled. We propose a *monotonicity likelihood ratio condition* that guarantees the existence and uniqueness of an "optimal" preorder on the input space, in the sense that it is optimal for any bipartite ranking subproblem, considering all possible pairs of labels. In particular, the regression function is proved to define an optimal ranking rule in this setting, highlighting the connection between K-partite ranking and ordinal regression. We show that the notion of ROC manifold/surface and its summary, the *volume under the* ROC *surface* (VUS), then provides quantitative criteria for evaluating ranking accuracy in the ordinal setup: under the afore mentioned monotonicity likelihood ratio condition, scoring functions whose ROC surface is as high as possible everywhere exactly coincide with those forming the optimal set (*i.e.* the set of scoring functions that are optimal for all bipartite subproblems, defined with no reference to the notions of ROC surface and VUS). Conversely, we prove that the existence of a

scoring function with such a dominating ROC surface implies that the monotonicity likelihood ratio condition is fulfilled.

Before presenting how to build a scoring function in order to solve the ranking problem (see part II), the purpose of the next chapter is to estimate the ROC surface using a labeled data set. Moreover, we show how to build accurate confidence regions using bootstrap procedure.

1.4 Proofs

Proof of Proposition 1.1.1

The assertions (3) $\Rightarrow$ (2), (2) $\Rightarrow$ (4) and (2) $\Rightarrow$ (5) are straightforward.

(1) $\Rightarrow$ (3) Recall that $\eta(x) = \sum_{k=1}^{K} k \cdot \eta_k(x)$. Our goal is to establish that: $\forall (x, x') \in \mathcal{X}^2$,

$$\Phi_{k,l}(x) < \Phi_{k,l}(x') \Rightarrow \eta(x) < \eta(x').$$

The proof is based on the next lemma.

Lemma 1.4.1. *Suppose Assumption 1 is satisfied. Let $(x, x') \in \mathcal{X}^2$. If there exists $1 \leq l < k \leq K$ such that $0 < \Phi_{k,l}(x) < \Phi_{k,l}(x')$, then for all $j \in \{1, ..., K\}$, we have*

$$\sum_{i=j}^{K} \eta_i(x) \leq \sum_{i=j}^{K} \eta_i(x'). \tag{1.5}$$

Additionally, a strict version of inequality (1.5) holds true when $j = l + 1$.

Proof. Let $(x, x') \in \mathcal{X}^2$ and $1 \leq l < k \leq K$ be such that $\Phi_{k,l}(x) < \Phi_{k,l}(x')$

Combining $\Phi_{k,l}(x) = \dfrac{p_l \eta_k(x)}{p_k \eta_l(x)}$ and $\eta_l(x) = 1 - \sum_{i \neq l} \eta_i(x)$, we clearly have

$$\eta_k(x) - \eta_k(x) \sum_{i \neq l} \eta_i(x') < \eta_k(x') - \eta_k(x') \sum_{i \neq l} \eta_i(x),$$

and, by virtue of Assumption 1, for $1 \leq j \leq m \leq K$:

$$\begin{aligned} \eta_m(x) &\leq \eta_m(x') + \sum_{i<j-1} \left\{ \eta_m(x)\eta_i(x') - \eta_m(x')\eta_i(x) \right\} \qquad (1.6) \\ &\quad + \sum_{i>j-1} \left\{ \eta_m(x)\eta_i(x') - \eta_m(x')\eta_i(x) \right\}, \\ &\leq \eta_m(x') + \sum_{i>j-1} \left\{ \eta_m(x)\eta_i(x') - \eta_m(x')\eta_i(x) \right\}. \qquad (1.7) \end{aligned}$$

Summing up, term-by-term, inequalities (1.6) for $m = j, \ldots, K$, one gets that

$$\sum_{m=j}^{K} \eta_m(x) \leq \sum_{m=j}^{K} \eta_m(x') + \sum_{m=j}^{K} \sum_{i=j}^{K} \left\{ \eta_m(x)\eta_i(x') - \eta_m(x')\eta_i(x) \right\}.$$

The proof is finished by noticing that the sum on the right hand side of the inequality above is equal to 0. □ □

The desired result is established by summing up the inequalities (1.5) stated in Lemma 1.4.1 for $j = 1, \ldots, K$.

(4) ⇒ (2) Using the fact that $\Phi_{k,l}(x) = \prod_{j=l}^{k-1} \Phi_{j+1,j}(x)$, immediatly gives the result.

(5) ⇒ (1) We call $\Psi_{k,l}$ the nondecreasing function such that $\Phi_{k,l}(x) = \Psi_{k,l}(s^*(x))$. Let $(k,l) \in \{1,\ldots,K\}^2$ s.t. $l < k$ and x, x' in $\mathcal{X}_k \cap \mathcal{X}_l$. Suppose that $\Phi_{k+1,k}(x) < \Phi_{k+1,k}(x')$. The functions $\Psi_{k,l}^{-1}$ are nondecreasing, just like the $\Psi_{k,l}$'s. The equality $\Phi_{l+1,l}(x) = \Psi_{l+1,l}(\Psi_{k+1,k}^{-1} \circ \Phi_{k+1,k}(x))$ on $\mathcal{X}_k \cap \mathcal{X}_l$ leads to the result.

Proof of Proposition 1.1.2

Notice first that it is actually sufficient to prove that $\mathcal{X}_{k-1} \cap \mathcal{X}_{k+1} \subset \mathcal{X}_k$ for all $k \in \{2, \ldots, K-1\}$. Let $1 < k < K$ and suppose that $\mathcal{X}_{k-1} \cap \mathcal{X}_k \neq \emptyset$ (the inclusion is immediate otherwise). Consider $x \in \mathcal{X}_{k-1} \cap \bar{\mathcal{X}}_k \cap \mathcal{X}_{k+1}$, where $\bar{\mathcal{X}}_k = \mathcal{X} \setminus \mathcal{X}_k$. We thus have: $\Phi_{k,k-1}(x) = 0$ and $\Phi_{k+1,k}(x) = +\infty$. Hence, for any $x' \in \mathcal{X}_k$, we have: $0 = \Phi_{k,k-1}(x) \leq \Phi_{k,k-1}(x')$ and $\Phi_{k+1,k}(x') \leq \Phi_{k+1,k}(x) = +\infty$. Assumption 1 implies that both inequalities are actually equalities, which is in contradiction with the fact that $x' \in \mathcal{X}_k$.

Proof of Proposition 1.2.1

This results from the change of parameters: $\alpha = F_{s,1}(t_1)$ and $\gamma = 1 - F_{s,3}(t_2)$.

Proof of Lemma 1.2.2

(1) ⇒ (2) If $\mathrm{ROC}(s,\alpha,\gamma) > 0$ using Proposition 1.2.1, we get $t_1 < t_2$, $\alpha = F_{s,1}(t_1)$ and $\gamma = 1 - F_{s,3}(t_2)$. Using the definition of the ROC curve and $t_1 < t_2$, we have $\mathrm{ROC}_{\phi_1,\phi_3}(s,1-\gamma) = 1-\mathrm{F}_{s,3}(t_1) > 1-\mathrm{F}_{s,3}(t_2) = \gamma$. (2) ⇒ (1) If $\mathrm{ROC}_{\phi_1,\phi_3}(s,1-\gamma) > \gamma$ then $1 - F_{s,3}(t_1) > 1 - F_{s,3}(t_2)$ so $t_1 < t_2$, and $F_{s,2}(t_2) - F_{s,2}(t_1) > 0$. This yields the desired result.

Proof of Proposition 1.2.3

We first prove assertion (1). If $\alpha = 0$ then $\mathrm{ROC}(s,0,\gamma) = \mathrm{F}_{s,2} \circ \mathrm{F}_{s,3}^{-1}(1-\gamma) = 1-\mathrm{ROC}_{\phi_3,\phi_2}(s,\gamma)$. Now, it easy to see that the curve $\{(\gamma, \mathrm{ROC}_{\phi_3,\phi_2}(s,\gamma))\}$ coincides with $\{(\beta, \mathrm{ROC}_{\phi_2,\phi_3}(s,\beta))\}$.

The proof of assertion (2) relies on the equality $F_{s,2} \circ F_{s,1}^{-1}(\alpha) = \mathbb{E}[\mathbb{I}\{s(X) < F_{s,1}^{-1}(\alpha)\}|Y=2] = \mathbb{E}[\mathbb{I}\{F_{s,1}(s(X)) < (\alpha)\}|Y=2]$. Using $F_{s,1}(s(X)) = \mathbb{E}[\mathbb{I}\{s(X') < s(X)\}|Y'=3]$ combined with the equality $\{(x,x') \in \mathcal{X}^2 | s(x') < s(x)\} = \{(x,x') \in \mathcal{X}^2 | s \circ T(X') < s \circ T(X)\}$ since T is strictly increasing, we obtain the desired result.

Turning to assertion (3), observe that if (α, γ) is such that $(F_{s,1}^{-1}(\alpha), F_{s,3}^{-1}(1-\gamma))$ belongs to $\{u, dF_{s,2}/dF_{s,1}(u) < \infty\} \times \{u, dF_{s,2}/dF_{s,3}(u) < \infty\}$ then

$$\frac{\partial}{\partial \alpha}\mathrm{ROC}(\mathrm{s}, \alpha, \gamma) = \frac{-\mathrm{dF_{s,2}}}{\mathrm{dF_{s,1}}}(\mathrm{F_{s,1}^{-1}}(\alpha)) = \psi_{21}(\mathrm{T}(\mathrm{F_{s,1}^{-1}}(\alpha)))$$

and

$$\frac{\partial}{\partial \alpha}\mathrm{ROC}(\mathrm{s}, \alpha, \gamma) = \frac{-\mathrm{dF_{s,2}}}{\mathrm{dF_{s,3}}}(\mathrm{F_{s,3}^{-1}}(1-\gamma)) = \frac{1}{\psi_{32}(\mathrm{T}(\mathrm{F_{s,1}^{-1}}(\alpha)))},$$

which are both (strictly) decreasing functions if $dF_{s,2}/dF_{s,1}(u)$ and $dF_{s,3}/dF_{s,2}(u)$ are both (strictly) increasing transforms of a certain function $T(u)$.

The derivatives calculated in (3) permit to establish assertion (4). Assertion (5) is straightforward.

Proof of Theorem 1.2.4

Let $(s, s^*) \in \mathcal{S} \times \mathcal{S}^*$. Since, in particular, the scoring function s^* belongs to the set $\mathcal{S}^*_{1,3}$, we have $\mathrm{ROC_{F_1,F_3}}(\mathrm{s}^*, 1-\alpha) \geq \mathrm{ROC_{F_1,F_3}}(\mathrm{s}, 1-\alpha)$ for all $\alpha \in [0,1]$. Hence, as the desired bound obviously holds true on the set $\{(\alpha, \gamma) :\ \gamma > \mathrm{ROC_{F_1,F_3}}(\mathrm{s}^*, 1-\alpha)\} \subset \{(\alpha, \gamma) :\ \gamma > \mathrm{ROC_{F_1,F_3}}(\mathrm{s}, 1-\alpha)\}$, we place ourselves on the complementary set $\{(\alpha, \gamma) :\ \gamma \leq \mathrm{ROC_{F_1,F_3}}(\mathrm{s}^*, 1-\alpha)\}$, on which we have

$$\begin{aligned}\mathrm{ROC}(\mathrm{s}^*, \alpha, \gamma) - \mathrm{ROC}(\mathrm{s}, \alpha, \gamma) \leq (\mathrm{ROC_{f_1,f_2}}(\mathrm{s}^*, 1-\alpha) - \mathrm{ROC_{f_1,f_2}}(\mathrm{s}, 1-\alpha)) + \\ (\mathrm{ROC_{f_3,f_2}}(\mathrm{s}, \gamma) - \mathrm{ROC_{f_3,f_2}}(\mathrm{s}^*, \gamma))\,.\end{aligned}$$

The terms on the right hand side of the equation are both nonnegative, since s^* lies in $\mathcal{S}^*_{1,2}$ and $\mathcal{S}^*_{3,2}$ respectively (observing that, whatever the two distributions H and G on $\mathbb{R}$ and for any $s \in \mathcal{S}$ and $(\alpha, \beta) \in [0,1]^2$, we have: $\mathrm{ROC_{H,G}}(\mathrm{s}, \alpha) \leq \beta \Leftrightarrow \alpha \leq \mathrm{ROC_{G,H}}(\mathrm{s}, \beta)$). The first part of the result is thus established.

Suppose that there exists $s^* \in \mathcal{S}$ such that, for any $s \in \mathcal{S}$, we have: $\forall (\alpha, \gamma) \in [0,1]^2$,

$$\mathrm{ROC}(\mathrm{s}^*, \alpha, \gamma) \geq \mathrm{ROC}(\mathrm{s}, \alpha, \gamma). \tag{1.8}$$

Observe that, if $\gamma > \mathrm{ROC_{f_1,f_3}}(\mathrm{s}^*, 1-\alpha)$, this implies that $\gamma > \mathrm{ROC_{f_1,f_3}}(\mathrm{s}, 1-\alpha)$, whatever (α, γ). It then follows that $s^* \in \mathcal{S}^*_{1,3}$. Now the fact that s^* belongs to $\mathcal{S}^*_{1,2}$ (respectively, to $\mathcal{S}^*_{1,3}$) straightforwardly result from Eq. (1.8) with $\beta = 0$ (respectively, with $\alpha = 1$).

Proof of Proposition 1.2.5

We denote by $\bar{E} = \mathcal{X} \setminus E$ the complementary set of any subset $E \subset \mathcal{X}$ and set $m_1(x) = \mathbb{I}\{x \in \bar{R}^{*(1)}_{\alpha}\} - \mathbb{I}\{x \in \bar{R}^{(1)}_{s,\alpha}\}$ and $m_3(x) = \mathbb{I}\{x \in R^{*(3)}_{1-\gamma}\} - \mathbb{I}\{x \in R^{(3)}_{s,1-\gamma}\}$ for $\alpha \in [0,1]$. On the set $\{(\alpha, \gamma) :\ \gamma \leq \mathrm{ROC_{f_1,f_3}}(\mathrm{s}^*, 1-\alpha)\}$, we may then write:

$$\mathrm{ROC}(\mathrm{s}^*, \alpha, \gamma) - \mathrm{ROC}(\mathrm{s}, \alpha, \gamma) \leq -\mathbb{E}[\mathrm{m_1(X)}|\mathrm{Y}=2] - \mathbb{E}[\mathrm{m_3(X)}|\mathrm{Y}=2].$$

Considering the first ROC curve deficit, we have:

$$-\mathbb{E}[m_1(X)|Y=2] = -\frac{p_1}{p_2}\mathbb{E}\left[m_1(X)\frac{\eta_2(X)}{\eta_1(X)}|Y=1\right].$$

Then we add and subtract $\frac{\eta_3(x)}{\eta_1(x)} - \frac{1-Q^{(1)}(\eta_1,\alpha)}{Q^{(1)}(\eta_1,\alpha)}$, this leads to:

$$\begin{aligned}-\mathbb{E}\left[m_1(X)|Y=2\right] = -\frac{p_1}{p_2}\mathbb{E}\left[m_1(X)(\frac{\eta_2(X)+\eta_3(X)}{\eta_1(X)} + \frac{1-Q^{(1)}(\eta_1,\alpha)}{Q^{(1)}(\eta_1,\alpha)})|Y=1\right] \\ + \frac{p_1}{p_2}\mathbb{E}\left[m_1(x)\frac{\eta_3(X)}{\eta_1(X)}|Y=1\right].\end{aligned}$$

By definition of s^*, the second term on the right hand side of the equation above is equal to

$$\frac{p_3}{p_2}\mathbb{E}[m_1(X)|Y=3] = \mathrm{ROC}_{\mathrm{f_1},\mathrm{f_3}}(\mathrm{s},1-\alpha) - \mathrm{ROC}_{\mathrm{f_1},\mathrm{f_3}}(\mathrm{s}^*,1-\alpha),$$

while, for the first term, by removing the conditioning with respect to $Y=1$ and using then the definition of $Q^{(1)}(\eta_1,\alpha)$), we get:

$$\frac{1}{p_2 Q^{(1)}(\eta_1,\alpha)}\mathbb{E}\left[m_1(X)\left(\eta_1(X) - Q^{(1)}(\eta_1,\alpha)\right)\right] = \frac{1}{p_2}\mathbb{E}\left[\left|\eta_1(X) - Q^{(1)}(\eta_1,\alpha)\right| m_1(X)\right].$$

The first part of the desired bound follows from $A\Delta B = \bar{A}\Delta\bar{B}$. The other ROC curve difference can be handled the same way. This leads to the desired result.

Proof of Proposition 1.2.8

We have:

$$\begin{aligned}\mathrm{VUS(s)} &= \int_0^1\int_0^1 \mathrm{ROC}(\mathrm{s},\alpha,\gamma)\ \mathrm{d}\alpha\mathrm{d}\gamma = \int_0^1\int_0^{1-\gamma}(1-\alpha-\gamma)\mathrm{d}\alpha\mathrm{d}\gamma \\ &= \frac{1}{2}\int_0^1 (1-\gamma)^2 d\alpha d\gamma\ = 1/6,\end{aligned}$$

which establishes the first assertion, while the second one results from:

$$\mathrm{VUS(s)} = \int_0^1\int_0^1 \mathrm{d}\alpha\mathrm{d}\gamma\ = 1.$$

Proof of Proposition 1.2.10

Proof. We show the proposition for scoring function with continuous distributions i.e. $\mathbb{P}\{s(X)=s(X')\}=0$. In this case, we have to show that

$$\mathrm{VUS(s)} = \mathbb{P}\left\{\mathrm{s(X_1)} < \ldots < \mathrm{s(X_K)}|\mathrm{Y_1}=1,\ldots,\mathrm{Y_K}=\mathrm{K}\right\}.$$

We prove the formula by mathematical induction. For $K = 2$, it is obviously true, see [Clémençon *et al.*, 2008]. Assume that the formula holds for $K-1$. For K, we recall that the ROC surface can be define by $\mathrm{ROC}(\mathrm{s}, \mathbf{u}) = 1 - \mathrm{F}_{\mathrm{s,K}}((\mathrm{c}_{\mathrm{K}-1}))$ where $\mathbf{u} \in (0,1)^{K-1}$, $c_k = F_{s,k}^{-1}(u_k + F_{s,k}(c_{k-1}))$ for $k = 2, \ldots, K-1$ and $c_1 = F_{s,1}^{-1}(u_1)$, see [Li & Zhou, 2009]. We have

$$\begin{aligned}\mathrm{VUS(s)} &= \int_{[0,1]^{\mathrm{n}}} \mathrm{ROC(s, u)d\mathbf{u}} = \int_{[0,1]^{\mathrm{n}}} 1 - \mathrm{F_{s,K}}(\mathrm{F}_{\mathrm{s,K-1}}^{-1}(\mathrm{u_{K-1}} + \mathrm{F_{s,K-1}}(\mathrm{c_{K-2}})))\mathrm{d\mathbf{u}} \\ &= \int_{[0,1]^n} \mathbb{E}[\mathbb{I}\{s(X_K) \geq F_{s,K-1}^{-1}(u_{K-1} + F_{s,K-1}(c_{K-2})\}|Y_K = K)]d\mathbf{u}\end{aligned}$$

By Fubini

$$\mathrm{VUS(s)} = \int_{[0,1]^{\mathrm{n-1}}} \mathbb{E}\left[\mathbb{E}[\mathbb{I}\{\mathrm{s(X_K)} \geq \mathrm{F}_{\mathrm{s,K-1}}^{-1}(\mathrm{U} + \mathrm{F_{s,K-1}}(\mathrm{c_{K-2}})\}|\mathrm{Y_K} = \mathrm{K})\right] \mathrm{d\mathbf{u}}$$

Now we show that $F_{s,K-1}^{-1}(U + F_{s,K-1}(c_{K-2})) \overset{d}{=} (s(X_{K-1})|s(X_{K-1}) \geq c_{K-2}, Y = K-1)$ in distribution. Indeed, we have

$$\begin{aligned}&\mathbb{P}(F_{s,K-1}^{-1}(U + F_{s,K-1}(c_{K-2})) < t) \\ &= \mathbb{P}(s(X_{K-1}) < F_{s,K-1}^{-1}(F_{s,K-1}(t) - F_{s,K-1}(c_{K-2}))|Y = K-1) \\ &= \mathbb{P}(F_{s,K-1}(s(X_{K-1})) < F_{s,K-1}(t) - F_{s,K-1}(c_{K-2})|Y = K-1)\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}((s(X_{K-1}) < t|s(X_{K-1}) \geq c_{K-2}, Y = K-1)) \\ &= \mathbb{P}(c_{K-2} < s(X_{K-1}) < t|Y = K-1) \\ &= \mathbb{P}(F_{s,K-1}(c_{K-2}) < F_{s,K-1}(s(X_{K-1})) < F_{s,K-1}(t)|Y = K-1)\end{aligned}$$

As $F_{s,K-1}(s(X_{K-1}))$ follow a uniform distribution, we obtain the equality in distribution. So we can write

$$\mathrm{VUS(s)} = \int_{[0,1]^{\mathrm{n-1}}} \mathbb{E}\left[\mathbb{I}\{\mathrm{s(X_K)} \geq \mathrm{s(X_{K-1})} \geq \mathrm{c_{K-2}}\}|\mathrm{Y_{K-1}} = \mathrm{K-1}, \mathrm{Y_K} = \mathrm{K}\right] \mathrm{d\mathbf{u}}.$$

Using the equality for a number of classes of $K-1$ leads to the desired result. □

Proof of Proposition 1.2.11

The result simply follows from integration over $(\alpha, \gamma) \in [0,1]^2$ of the inequality stated in Theorem 1.2.4.

Proof of Theorem 1.2.12

Let $(s, s^*) \in \mathcal{S} \times \mathcal{S}^*$. Notice that, as $s^* \in \mathcal{S}^*_{1,3}$, we have $\{(\alpha, \gamma) : \ \gamma \leq \mathrm{ROC}_{\phi_1,\phi_3}(\mathrm{s}, 1-\alpha)\} \subset \{(\alpha, \gamma) : \ \gamma \leq \mathrm{ROC}_{\phi_1,\phi_3}(\mathrm{s}^*, 1-\alpha)\}$, so that

$$\begin{aligned}
\mathrm{ROC}^*(\alpha,\gamma) - \mathrm{ROC}(\mathrm{s},\alpha,\gamma) &\leq \{\mathrm{ROC}_{\phi_1,\phi_2}(\mathrm{s}^*, 1-\alpha) - \mathrm{ROC}_{\phi_3,\phi_2}(\mathrm{s}^*,\gamma) \\
&- (\mathrm{ROC}_{\phi_1,\phi_2}(\mathrm{s}, 1-\alpha) - \mathrm{ROC}_{\phi_3,\phi_2}(\mathrm{s},\gamma))_+\} \\
&\times \mathbb{I}\{\gamma \leq \mathrm{ROC}^*_{\phi_1,\phi_3}(1-\alpha)\} \\
&\leq \{\mathrm{ROC}_{\phi_1,\phi_2}(\mathrm{s}^*, 1-\alpha) - \mathrm{ROC}_{\phi_3,\mathrm{f}\phi_2}(\mathrm{s}^*,\gamma) \\
&- \mathrm{ROC}_{\phi_1,\phi_2}(\mathrm{s}, 1-\alpha) - \mathrm{ROC}_{\phi_3,\phi_2}(\mathrm{s},\gamma)\} \\
&\times \mathbb{I}\{\gamma \leq \mathrm{ROC}^*_{\phi_1,\phi_3}(1-\alpha)\} \\
&\leq (\mathrm{ROC}_{\phi_1,\phi_2}(\mathrm{s}^*, 1-\alpha) - \mathrm{ROC}_{\phi_1,\phi_2}(\mathrm{s}, 1-\alpha)) \\
&- (\mathrm{ROC}_{\phi_3,\phi_2}(\mathrm{s}^*,\gamma) - \mathrm{ROC}_{\phi_3,\phi_2}(\mathrm{s},\gamma))\,.
\end{aligned}$$

Integrating over $(\alpha, \gamma) \in [0,1]^2$ then yields the desired bound, using the fact that, for any $s \in \mathcal{S}_0$, $\int_{\gamma=0}^1 \mathrm{ROC}_{\phi_3,\phi_2}(\mathrm{s},\gamma)\mathrm{d}\gamma = 1 - \mathrm{AUC}_{\phi_2,\phi_3}(\mathrm{s})$.

Proof of Proposition 1.2.13

By virtue of Theorem 1.2.12, we have:

$$\mathrm{VUS}^* - \mathrm{VUS}(\widehat{\eta}) \leq \left(\mathrm{AUC}^*_{\phi_1,\phi_2} - \mathrm{AUC}_{\phi_1,\phi_2}(\widehat{\eta})\right) + \left(\mathrm{AUC}^*_{\phi_2,\phi_3} - \mathrm{AUC}_{\phi_2,\phi_3}(\widehat{\eta})\right).$$

Considering the first term on the right hand side of the equation above, we have:

$$\mathrm{AUC}^*_{\phi_1,\phi_2} - \mathrm{AUC}_{\phi_1,\phi_2}(\widehat{\eta}) = \frac{1}{2\mathrm{p}_1\mathrm{p}_2}\mathbb{E}[|\eta_1(\mathrm{X})\eta_2(\mathrm{X}') - \eta_1(\mathrm{X}')\eta_2(\mathrm{X})| \cdot \mathbb{I}\{(\mathrm{X},\mathrm{X}') \in \Gamma\}],$$

where

$$\Gamma = \{(x, x') \in \mathcal{X}^2 : \ \ (\eta(x) - \eta(x'))(\widehat{\eta}(x) - \widehat{\eta}(x')) < 0\}.$$

By using the triangular inequality and Lemma 1.4.1, one may establish that: $\forall (x, x') \in \mathcal{X}^2$, $\forall i \in \{1, 2, 3\}$,

$$|\eta_i(x) - \eta_i(x')| < |\eta(x) - \eta(x')|.$$

Then, we get:

$$\mathrm{AUC}^*_{\phi_1,\phi_2} - \mathrm{AUC}_{\phi_1,\phi_2}(\widehat{\eta}) \leq \frac{1}{2\mathrm{p}_1\mathrm{p}_2}\mathbb{E}[|\eta(\mathrm{X}) - \eta(\mathrm{X}')|\mathbb{I}\{(\mathrm{X},\mathrm{X}') \in \Gamma\}].$$

But, one may easily check that, if $(x, x') \in \Gamma$, then

$$|\eta(x) - \eta(x')| \leq |\eta(x) - \widehat{\eta}(x)| + |\eta(x') - \widehat{\eta}(x')|.$$

As the same argument can be applied to the second AUC difference, this gives the desired result.

1.5 Annex - Fast computation of the VUS in case of ties

It is the purpose of this annex to present an algorithm to compute the empirical VUS of a scoring function s that might have ties. First we recall the probabilistic version of the VUS for when there is k labels :

$$\mathrm{VUS(s)} = \sum_{\mathrm{u}\in\mathcal{B}^{(\mathrm{K})}} \frac{\mathbb{P}\{\cap_{\mathrm{k=1}}^{\mathrm{K-1}}\mathcal{E}^{\mathrm{u(k)}}(\mathrm{X_k},\mathrm{X_{k+1}})|\mathrm{Y_1}=1,\ldots,\mathrm{Y_K}=\mathrm{K}\}}{\prod_{\mathrm{i=1}}^{|\mathrm{g_u}|_1}(\mathrm{g_u(i)}+1)!},$$

where $\mathcal{B}^{(k)} = \{0,1\}^{k-1}$, $\mathcal{E}^0(X_k, X_{k+1})$ is the event "$s(X_k) < s(X_{k+1})$", $\mathcal{E}^1(X_k, X_{k+1})$ is the event "$s(X_k) = s(X_{k+1})$", g_u is the vector counting the number of consecutive 1 in u.

Let $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ a sample. We denote σ the permutation that order $\{s(x_1), \ldots, s(x_n)\}$ and $\sigma_1, \ldots, \sigma_M$ the unique values of this set. We introduce the notation

$$q(s, D, k, m) = \sum_{(i_1,\ldots,i_k)|s(i_1)\leq\ldots\leq s(i_1)\leq\sigma_m} \sum_{u\in\mathcal{B}^{(k)}} \frac{\mathbb{I}\{\cap_{k=1}^{K-1}\mathcal{E}^{u(k)}(X_k, X_{k+1})\}}{\prod_{i=1}^{|g_u|_1}(g_u(i)+1)!}.$$

Here $q(s, D, k, m)$ can be interpreted as the number of ordered k-tuples (with the correction when ties occur) for the m first unique values of the function s. We have $\mathrm{VUS(s)} = \frac{1}{\prod_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{n_k}} \mathrm{q(s, D, K, M)}$.

Computation of the empirical VUS

Input. Data sample $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, a scoring function s.

1. Sort $\mathcal{D}$ according to s to obtain a permutation σ.
2. Create the vector of values $\sigma_1, \ldots, \sigma_M$ and the corresponding number of labels with these values $(z(i,j))_{1\leq i\leq M, 1\leq j\leq K}$
3. For $i = 1$ to M do for $k = 1$ to K do for $j = 1$ to $k-1$ to k do $q(s, D, k) = q(s, D, k) + (\prod_l^k (z(i,j)) * q(s, D, j)/(k-j)!)$
4. $\widehat{\mathrm{VUS}}(s, \mathcal{D}) = \frac{1}{\prod_{k=1}^{K} n_k} q(s, D, K)$

Output. $\widehat{\mathrm{VUS}}(s, \mathcal{D})$

Obviously the computation of the order is the most costly part so this algorithm is in $O(n \log(n))$

Chapter 2

Confidence regions for the ROC surface via smoothed bootstrap

In this chapter, we consider that the scoring function is known as well as a labeled dataset $\mathcal{D}_n = \{(X_i, Y_i)\ i = 1, \ldots, n\}$ with $\mathcal{Y} = \{1, 2, 3\}$. The goal is to estimate the ROC surface described in the previous chapter and build confidence regions for the empirical ROC surface. These are important questions when the goal is to compare scoring functions and deciding which one is the best. Moreover, an important question is the confidence of such a decision.

The ROC curve has been widely used in a variety of applications such as signal processing, information retrieval (IR) and credit-risk screening, refer to [Fawcett, 2006]. In medicine for instance, it is used to evaluate diagnostic tests, which aim at discriminating the patients suffering from a given disease from the others by means of physiochemical measurements and/or the possible occurrence of certain symptoms, see [Pepe, 2003]. Nonparametric and semi-parametric estimation of the ROC curve has been investigated at length in [Hsieh & Turnbull, 1996]. The construction of pointwise confidence intervals for the ROC curve or for summary quantities such as the AUC is tackled in [Macskassy & Provost, 2004], [Hall *et al.*, 2004] and [Cortes & Mohri, 2004].

However, in the multipartite ranking case, more than two populations are involved and we have to estimate the ROC surface, see Chapter 1. Numerical issues related to the computation of empirical ROC surfaces have been documented in the medical literature, see [Nakas & Yiannoutsos, 2004] and [Li & Zhou, 2009]. In addition, statistical results related to the estimation of the ROC surface can be found in [Li & Zhou, 2009], extending those established in [Hsieh & Turnbull, 1996] in the bipartite context.

It is the major purpose of this chapter to show how to build confidence regions for the ROC surface, in a computationally feasible manner, using the bootstrap methodology. Since the multipartite ranking is a global task, all the ROC surface is important. So the problem of estimation is tackled from a functional angle. In such a framework, involving resampling in a path space, a smooth variant of the bootstrap procedure should be preferred to the naive bootstrap, see [Falk & Reiss, 1989]. The idea is to generate the bootstrapped data using a smooth version of the class distributions and then build a bootstrap ROC surface based on the data thus sampled. Theoretical results supporting this approach for building confidence regions and based on strong approximation results are given. Experiments are also carried

out to illustrate the performance of the methodology promoted in this chapter. We point out that this approach has been proposed and briefly described in the bipartite context at the NIPS 2008 conference, see [Bertail *et al.*, 2008].

The rest of the chapter is organized as follows. In section 2.1, we present a non parametric estimation of the ROC surface. The choice of an adequate metric on the ROC space is also considered. In section 2.2, the nonparametric estimator of the ROC surface is shown to be asymptotically Gaussian. The smoothed bootstrap algorithm is next described at length in section 2.2.2, together with the theoretical results establishing its asymptotic accuracy. Section 2.3 is devoted to displaying numerical results, illustrating the advantage of the smooth bootstrap technique on simulated data. Our methodology is also applied to a SWD dataset in order to compare the performance of two ranking algorithms. Technical details are postponed to section 2.5.

2.1 Estimation and distances

In this section, we tackle the issue of estimating the ROC surface of a given scoring function s, based on a pooled sample of $n \geq 1$ labeled data $\mathcal{D}_n = \{(X_i, Y_i) : \ 1 \leq i \leq n\}$, independent copies of a random pair (X, Y) where Y is drawn from the distribution $p_1\delta_1 + p_2\delta_2 + p_3\delta_3$, F_k being the conditional distribution of X given $Y = k$ and $p_k = \mathbb{P}\{Y = k\}$, with $k \in \{1,\ 2,\ 3\}$.

2.1.1 Metrics on the ROC space

Here, we propose several metrics to measure closeness in the ROC space. Let $\rho : [0,1]^2 \to \mathbb{R}$ be a measurable function and we denote by $\rho_y(.) = \rho(., y)$ (resp. by $\rho_x(.) = \rho(x, .)$) its restriction to the first (resp. second) coordinate for a fixed y (resp. x). We also denote by $\mathbb{D}([0,1]^2)$ the Skorohod space on the unit square, *i.e.* the set of functions $\rho : [0,1]^2 \to \mathbb{R}$ such that $\forall x \in (0,1), \forall y \in (0,1)\ \lim_{u \to x^+} \rho_y(u) = \rho(x, y)$, $\lim_{u \to y^+} \rho_x(u) = \rho(x, y)$, $\limsup_{u \to x^-} \rho_y(u) < \infty$ and $\limsup_{u \to y^-} \rho_x(u) < \infty$. Viewing the ROC space as a subset of $\mathbb{D}([0,1]^2)$, the standard metrics are the sup norm $\|.\|_\infty$ and the *Hausdorff distance* d_H. The Hausdorff distance between two elements ρ_1 and ρ_2 of $\mathbb{D}([0,1]^2)$ is the following quantity :

$$d_H(\rho_1, \rho_2) = \max\{ \sup_{t \in [0,1]^2} \inf_{u \in [0,1]^2} |\rho_1(t) - \rho_2(u)|,\ \sup_{u \in [0,1]^2} \inf_{t \in [0,1]^2} |\rho_1(t) - \rho_2(u)| \}.$$

If the former choice appears natural when considering smooth functions, the latter is more pertinent for analyzing stepwise graphs, such as the empirical ROC surfaces. Indeed, equipped with this metric, two piecewise constant ROC surfaces may be close to each other, even if their jumps do not exactly match. However, providing a theoretical basis in the case of Hausdorff distance is very challenging and computing in practice the distance is not straightforward. As shall be seen below, asymptotic arguments for grounding the bootstrapping of the empirical ROC surface fluctuations, when measured in terms of sup norm $\|.\|_\infty$, are possible. However, given the

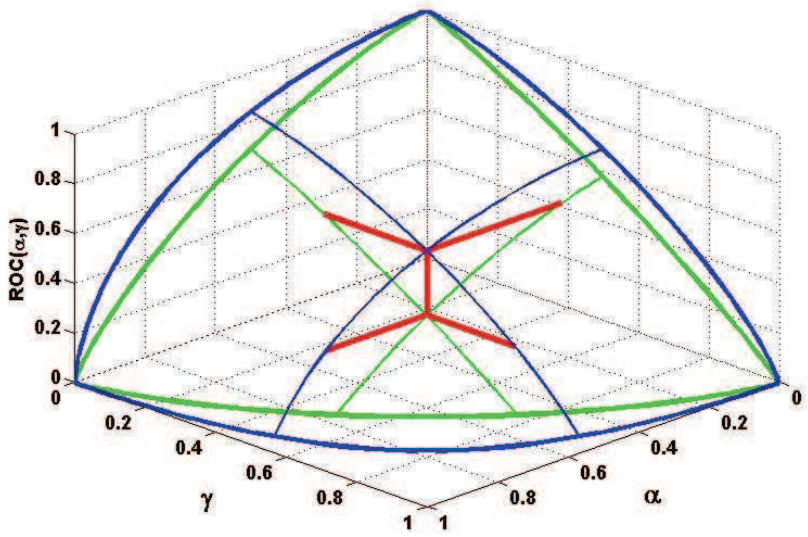


Figure 2.1: The d_B distance of two ROC surfaces at point $(\alpha, \gamma) = (1/2, 1/2)$ is the minimum among the lengths of the red segments.

geometry of empirical ROC surfaces this metric is not convenient for our purpose and possibly produce very wide, and thus non informative, confidence regions. This can be explain by the possible occurrence of nearly vertical tangent planes of the ROC surface near the origin $(0,\ 0)$. Notice that the ROC surface is a coordinatewise non-decreasing function and denote by $\mathcal{M}$ the set of functions $\rho : [0,1]^2 \to \mathbb{R}$ such that ρ_x and ρ_y are monotone functions, for all $x, y \in [0,1]$. Using the monotonicity of the ROC surfaces, we shall consider the closely related pseudo metric d_B defined as follows :

$$\forall (\rho_1, \rho_2) \in \mathbb{D}([0,1]^2)^2 \cap \mathcal{M}^2, \quad d_B(\rho_1, \rho_2) = \sup_{t_1, t_2 \in [0,1]} d_B(\rho_1, \rho_2, t_1, t_2),$$

where $d_B(\rho_1, \rho_2, t_1, t_2)$ denotes the minimum among these five quantities,

$$|\rho_1(t_1,t_2) - \rho_2(t_1,t_2)|, |\rho^{-1}_{1,x=t_1}(\rho_2(t_1,t_2)) - t_2|, |\rho^{-1}_{2,x=t_1}(\rho_1(t_1,t_2)) - t_2|,$$
$$|\rho^{-1}_{2,y=t_2}(\rho_1(t_1,t_2)) - t_1|, |\rho^{-1}_{1,y=t_2}(\rho_2(t_1,t_2)) - t_1|,$$

where $\rho^{-1}_{i,x=t_1}$ (resp $\rho^{-1}_{i,y=t_2}$) denotes the generalized inverse of the function $\rho_{i,t_1}(.) = \rho_i(t_1,.)$ (resp $\rho_{i,t_2}(.) = \rho_i(.,t_2)$). These quantities are depicted in Fig. 1. Observe that we clearly have:

$$d_H(\rho_1, \rho_2) \leq d_B(\rho_1, \rho_2) \leq \|\rho_1 - \rho_2\|_\infty.$$

The rationale behind this metric is to symmetrize the sup norm in order to weaken the impact of possible nearly vertical tangent plans of the ROC surface. This way, d_B

permits to build confidence regions of reasonable size, well adapted to the stepwise shape of empirical ROC surfaces.

2.1.2 Statistical estimation of the ROC surface

As proposed by [Li & Zhou, 2009], we estimate the ROC surface of a scoring function s by replacing in Eq. (1.1) the (unknown) distribution functions $F_{s,k}$, $1 \leq k \leq 3$, by their statistical counterparts based on the data sample $\mathcal{D}_n$. This yields the estimator given by:

$$\forall (\alpha, \gamma) \in [0,1]^2, \widehat{\mathrm{ROC}}_s(\alpha, \gamma) = \left(\widehat{F}_{s,2} \circ \widehat{F}_{s,3}^{-1}(1-\gamma) - \widehat{F}_{s,2} \circ \widehat{F}_{s,1}^{-1}(\alpha) \right)_+$$

with : $\widehat{F}_{s,k}(t) = (1/n_k) \sum_{i=1}^n \mathbb{I}\{Y_i = k\} \mathbb{I}\{t - s(X_i) \geq 0\}$, where $n_k = \sum_{i=1}^n \mathbb{I}\{Y_i = k\}$ is the random number of instances with label k among the pooled sample. In order to obtain smoothed version $\tilde{F}_{s,k}$ of the cdf's, a typical choice consists in substituting the indicator function by a smoothing function $K_h : \mathbb{R}^d \to \mathbb{R}$ such that $K_h(u) = h^{-1} K(h^{-1} u)$ where K is Parzen-Rosenblatt kernel bounded away from 0 on a neighborhood of 0 in $\mathbb{R}$ and $\int K(v) dv = 1$ and K is bounded, see [Silverman & Young, 1987]. The parameter $h > 0$ is called the smoothing bandwidth.

Notice that, in practice, to draw the empirical ROC surface using the dataset $\mathcal{D}_n$, it suffices to compute $\alpha(t_1, t_2) = \hat{F}_{s,1}(t_1)$, $\beta(t_1, t_2) = \hat{F}_{s,2}(t_2) - \hat{F}_{s,2}(t_1)$ and $\gamma(t_1, t_2) = 1 - \hat{F}_{s,3}(t_2)$ for all possible threshold values (t_1, t_2) in $\{s(X_i) : \quad i = 1, \ldots, n\}$ with $t_1 \leq t_2$, and plot the related piecewise constant surface in the ROC cube $[0,1]^3$, see Fig. 2.3 and 2.4 for example.

2.2 Assessment

In this section, we present the theoretical results of this chapter. First, we state an asymptotic theorem, showing the consistency and a strong approximation of this estimator. Next, we introduce a procedure to build confidence regions for the ROC surface using a smooth estimation of the distribution functions.

2.2.1 Asymptotic normality

Here, we study asymptotic properties of the empirical ROC surface, completing the results proved in [Li & Zhou, 2009]. The sup norm is used here to evaluate the distance between two surfaces. We also present a strong approximation of the fluctuation process

$$r_n(\alpha, \gamma) = \sqrt{n} \left(\widehat{\mathrm{ROC}}_s(\alpha, \gamma) - \mathrm{ROC}_s(\alpha, \gamma) \right), (\alpha, \gamma) \in [0,1]^2.$$

For this purpose, the following technical assumptions are required.

$\mathbf{A_1}$ The intersection of the tangent planes of the ROC surface with facets of the positive orthant have bounded slopes:

$$\sup_{\alpha\in[0,1]} \frac{f_{s,2}(F_{s,1}^{-1}(\alpha))}{f_{s,1}(F_{s,1}^{-1}(\alpha))} < \infty, \ \sup_{\gamma\in[0,1]} \frac{f_{s,2}(F_{s,3}^{-1}(1-\gamma))}{f_{s,3}(F_{s,3}^{-1}(1-\gamma))} < \infty.$$

$\mathbf{A_2}$ The cdf $F_{s,3}$ et $F_{s,1}$ are twice differentiable on $[0,\ 1]$ and

$$\forall(\alpha,\gamma) \in [0,1]^2, f_{s,1}(\alpha) > 0 \text{ and } f_{s,3}(\gamma) > 0.$$

In addition, there exists $\xi > 0$ such that

$$\sup_{\gamma\in[0,\ 1]} \gamma(1-\gamma)\frac{f'_{s,3}(F_{s,3}^{-1}(1-\gamma))}{f^2_{s,3}(F_{s,3}^{-1}(1-\gamma))} \leq \xi < \infty, \ \sup_{\alpha\in[0,1]} \alpha(1-\alpha)\frac{f'_{s,1}(F_{s,1}^{-1}(\alpha))}{f^2_{s,1}(F_{s,1}^{-1}(\alpha))} \leq \xi < \infty.$$

These hypotheses are standard in strong approximation theory, see [Csorgo & Revesz, 1981]. Equipped with these conditions, we can now state a limit theorem for empirical ROC surface establishing the consistency in sup norm and the asymptotic normality of the fluctuation process.

Theorem 2.2.1. *Under assumptions* $\mathbf{A_1} - \mathbf{A_2}$*, the following assertions hold true.*

(*i*) *(Consistency) With probability one, we have:*

$$\sup_{0\leq\alpha,\gamma\leq 1} |\widehat{ROC}(s,\alpha,\gamma) - ROC(s,\alpha,\gamma)| \to 0 \text{ when } n \to \infty.$$

(*ii*) *(Asymptotic accuracy) There exists a sequence of three independent Brownian bridges* $\{(B_1^{(n)}(t), B_2^{(n)}(t), B_3^{(n)}(t)), 0 \leq t \leq 1\}$ *with* $n \geq 1$*, such that we almost surely have, uniformly over* $[0,\ 1]^2$*,*

$$r_n(\alpha,\gamma) = \mathbb{I}\{\mathcal{I}_s\}z^{(n)}(\alpha,\gamma) + O((\log\log n)^{\rho_1(\xi)}\log^{\rho_2(\xi)} n/\sqrt{n}),$$

where

$$z^{(n)}(\alpha,\gamma) = \frac{1}{\sqrt{p_2}}B_1^{(n)}(F_{s,2}(F_{s,3}^{-1}(1-\gamma))) + \frac{1}{\sqrt{p_3}}\frac{f_{s,2}(F_{s,3}^{-1}(1-\gamma))}{f_{s,3}(F_{s,3}^{-1}(1-\gamma))}B_2^{(n)}(\gamma)$$
$$- \frac{1}{\sqrt{p_2}}B_1^{(n)}(F_{s,2}(F_{s,1}^{-1}(\alpha))) + \frac{1}{\sqrt{p_1}}\frac{f_{s,2}(F_{s,1}^{-1}(\alpha))}{f_{s,1}(F_{s,1}^{-1}(\alpha))}B_3^{(n)}(\alpha), \quad (2.1)$$

and

$$\begin{cases} \rho_1(\xi) = 0, \rho_2(\xi) = 1 \text{ if } \xi < 1 \\ \rho_1(\xi) = 0, \rho_2(\xi) = 2 \text{ if } \xi = 1 \\ \rho_1(\xi) = \gamma, \rho_2(\xi) = \xi - 1 + \varepsilon \text{ if } \xi > 1 \end{cases}.$$

We point out that the result established in [Li & Zhou, 2009] holds in distribution only and, incidentally, that the indicator function is missing in the description of the fluctuation process (it is not taken in consideration in their proof).

2.2.2 Smooth bootstrap

From a computational perspective, the asymptotic results stated in the previous section can hardly be used to build asymptotic confidence regions for ROC surface. Indeed, it would require to simulate Brownian bridges. In addition, the densities $f_{s,k}$'s are involved in Eq. (2.1). Accurate estimates of the class densities are thus necessarily required to build confidence regions this way. In this subsection, we explain how to construct confidence regions for the ROC surface via the bootstrap approach, originally introduced in [Efron, 1979]. The idea to consider, as an estimate of the law of the fluctuation process $r_n = \{r_n(\alpha,\gamma)\}_{(\alpha,\gamma)\in[0,1]^2}$, the conditional law given $\mathcal{D}_n$ of the bootstrapped fluctuation process

$$\tilde{r}_n(\alpha,\gamma) = \sqrt{n}(\widetilde{\mathrm{ROC}}_s(\alpha,\gamma) - \widehat{\mathrm{ROC}}_s(\alpha,\gamma)), ((\alpha,\gamma)) \in [0,1]^2, \tag{2.2}$$

where $\widetilde{\mathrm{ROC}}_s$ is the ROC surface corresponding to a sample $\tilde{\mathcal{D}}_n = \{(\tilde{Z}_i, \tilde{Y}_i)\}_{1\leq i\leq n}$ of i.i.d. random pairs with common distribution $\tilde{\mathcal{P}}_n$ close to the empirical distribution $\mathcal{P}_n$ of the $(s(X_i), Y_i)$'s. We also consider

$$\tilde{d}_n = \sqrt{n} d_B(\widetilde{\mathrm{ROC}}_s, \widehat{\mathrm{ROC}}_s),$$

whose random fluctuations given $\mathcal{D}_n$ are expected to mimic those of the quantity $d_n = \sqrt{n} d_B(\widehat{\mathrm{ROC}}_s, \mathrm{ROC}_s)$. Notice that the target of the bootstrap procedure is here a distribution on a path space, the ROC space being viewed as a subspace of $\mathbb{D}([0,1]^2)$ equipped with either $\|.\|_\infty$ or else $d_B(.,.)$. Since the estimate of the ROC surface involves the quantile processes $\{\widehat{F}_{s,1}^{-1}(\alpha)\}$ and $\{\widehat{F}_{s,3}^{-1}(\gamma)\}$, choosing the empirical distribution as resampling distribution, *i.e.* $\tilde{\mathcal{P}}_n = \hat{\mathcal{P}}_n$, may not be appropriate. In such situation, it is well-known that the smooth bootstrap, that consists in choosing a smooth version of the empirical distribution, improves over the naive bootstrap, see [Falk & Reiss, 1989]. We describe below the smooth bootstrap algorithm for a building confidence region at level $1-\varepsilon$ in the ROC space from sampling data $\mathcal{D}_n(s) = \{(s(X_i), Y_i) : \ 1 \leq i \leq n\}$.

We now investigate the asymptotic properties of the bootstrap method above. We point out that the result is of functional nature because in applications the whole estimation of the ROC surface, or some part of it at least, is what we need. In the sequel, we assume that the kernel $K(.)$ used in the smoothing step is Gaussian or of the form $\mathbb{I}\{u \in [-1;+1]\}$.

Theorem 2.2.2. *Suppose that Theorem 2.2.1's assumptions are fulfilled. Assume that smoothed version of the cdf's $\tilde{F}_{s,1}, \tilde{F}_{s,2}, \tilde{F}_{s,3}$ are computed using a scaled kernel $K_{h_n}(u)$ with $h_n \downarrow 0$ as $n \to \infty$ in a way that $nh_n^3 \to \infty$ and $nh_n^5 \log^2(n) \to 0$. Then, the bootstrap distribution estimates output by Algorithm 1 are such that:*

$$\sup_{t\in\mathbb{R}} |\mathbb{P}^*\{\|r_n^*\|_\infty \leq t\} - \mathbb{P}\{\|r_n\|_\infty \leq t\}| = O_{\mathbb{P}}\left(\frac{log(h_n^{-1})}{\sqrt{nh_n}}\right).$$

Algorithm 1

INPUT: dataset $\mathcal{D}_n(s) = \{s(X_i), Y_i : \ 1 \leq i \leq n\}$, level of confidence ε.
Compute the ROC surface estimate :

$$\forall (\alpha, \gamma) \in [0,1]^2, \widehat{\mathrm{ROC}}_s(\alpha, \gamma) = \left(\hat{F}_{s,2} \circ \hat{F}_{s,3}^{-1}(1-\gamma) - \hat{F}_{s,2} \circ \hat{F}_{s,1}^{-1}(\alpha) \right)_+$$

Draw a bootstrap sample $\tilde{\mathcal{D}}_n$ from the smooth distribution estimate :

$$\tilde{\mathcal{P}}(dz, y) = \frac{n_1}{n} \mathbb{I}\{Y = 1\} \tilde{F}_{s,1}(dz) + \frac{n_2}{n} \mathbb{I}\{Y = 2\} \tilde{F}_{s,2}(dz) + \frac{n_3}{n} \mathbb{I}\{Y = 3\} \tilde{F}_{s,3}(dz)$$

Compute the bootstrap version of the empirical class cdf estimates $F_{s,1}^*, F_{s,2}^*, F_{s,3}^*$ based on $\tilde{\mathcal{D}}_n$.
Return the confidence band at level $1-\varepsilon$ defined by the ball of center $\widehat{\mathrm{ROC}}$ and radius δ_ε where δ_ε is defined by $\mathbb{P}^*\{\|r_n^*\|_\infty \leq \delta_\varepsilon\} = 1-\varepsilon$ in the case of the sup norm or by $\mathbb{P}^*\{\tilde{d}_n \leq \delta_\varepsilon\} = 1-\varepsilon$ when considering the distance d_B, denoting by $\mathbb{P}^*$ the conditional probability given the original data $\mathcal{D}_n(s)$.
OUTPUT: Bootstrap confidence region at level 1-ε.

To obtain a similar Theorem for the confidence regions using the distance d_B, we need to add the same assumption as $\mathbf{A_2}$ for the cdf $F_{s,2}$.

$\mathbf{A_3}$ The cdf $F_{s,2}$ is twice differentiable on $[0,\ 1]$ and

$$\forall \beta \in [0,1], \quad f_{s,2}(\beta) > 0.$$

In addition, there exists $\xi > 0$ such that

$$\sup_{\beta \in [0,\ 1]} \beta(1-\beta) \frac{f'_{s,2}(F_{s,2}^{-1}(\beta))}{f_{s,2}^2(F_{s,2}^{-1}(\beta))} \leq \xi < \infty.$$

Theorem 2.2.3. *Suppose that the assumptions* $\mathbf{A_1}$, $\mathbf{A_2}$ *and* $\mathbf{A_3}$ *hold. Assume that smoothed version of the cdf's* $\tilde{F}_{s,1}, \tilde{F}_{s,2}, \tilde{F}_{s,3}$ *are computed using a scaled kernel* $K_{h_n}(u)$ *with* $h_n \downarrow 0$ *as* $n \to \infty$ *in a way that* $nh_n^3 \to \infty$ *and* $nh_n^5 \log^2(n) \to 0$. *Then, the bootstrap distribution estimates output by Algorithm 1 are such that:*

$$\sup_{t \in \mathbb{R}} |\mathbb{P}^*\{\tilde{d}_n \leq t\} - \mathbb{P}\{d_n \leq t\}| = O_{\mathbb{P}}\left(\frac{log(h_n^{-1})}{\sqrt{nh_n}} \right).$$

Picking the bandwidth h_n of order $1/\log(n^{2+\varepsilon})n^{1/5}$ with $\varepsilon > 0$ thus leads to an approximation error of order $\left(n^{-2/5}\right)$ up to a logarithm factor for the bootstrap distribution estimate. This rate is slower than the one of the Gaussian approximation given in Theorem 2.2.1(ii), the ROC surface bootstrap algorithm is however very appealing from computational angle. The construction of Gaussian confidence bands

from estimates of the densities $f_{s,1}$, $f_{s,2}$ and $f_{s,3}$ and simulated Brownian bridges is indeed very challenging to implement. The rate reached by smoothed bootstrap distribution is nevertheless a great improvement, compared to the naive bootstrap approach which order is $O_{\mathbb{P}}\left(n^{-1/4}\right)$. In practice, the true smooth bootstrap distribution can not be evaluated exactly and we have to resort to a Monte Carlo procedure to approximate numerically. It consists in drawing B bootstrap independent samples of size n from the smoothed distribution function $\tilde{\mathcal{P}}(dz, y)$. A practical way to draw from the smooth bootstrap distribution is to draw a sample from the empirical distribution function and then perturbing each data by an independent centered Gaussian random variable of variance h^2, see [Silverman & Green, 1986] for more details.

2.3 Numerical experiments

The purpose of this section is to investigate the performance of the smooth bootstrap algorithm proposed previously, in comparison with that of the naive bootstrap in particular, on several examples. The results are expressed in terms of coverage probability, *i.e.* the frequency at which the true ROC surface falls into the confidence region.

2.3.1 Simulated data

We consider first a toy example to evaluate the benefit of the smooth bootstrap. The impact of the metric, either the sup norm $\|.\|_\infty$ or else the distance d_B, is also quantified. The simple trinormal model has been simulated:

$$Y = \begin{cases} 1 & \text{if } \beta_0 + \beta_1 X + \varepsilon < 0.4 \\ 2 & \text{if } 0.4 \leq \beta_0 + \beta_1 X + \varepsilon < 1.6 \\ 3 & \text{if } 1.6 \leq \beta_0 + \beta_1 X + \varepsilon \end{cases},$$

where ε and X are independent standard $\mathcal{N}(0,1)$ r.v.'s. In this case, one may easily check that $s(x) = x$ is an optimal scoring function. We choose here $\beta_0 = \beta_1 = 1$, $n = B = 1000$ and $1 - \varepsilon = 0.95$ the targeted coverage probability. In addition, we take $h = n^{-1/5}$ as smoothing bandwidth. The estimated coverage probabilities are obtained over 2000 replications of the procedure and are reported in Table 2.1. We also depict confidence regions in Fig. 2.2(a) and (b).

In both experiments, the smooth bootstrap improves significantly the quality of the confidence region. As it has been already observed in the case of ROC curves (see [Hall *et al.*, 2004] or [Bertail *et al.*, 2008]), the naive bootstrap generally provides confidence regions that are much too small. We also point out that the symmetrized distance d_B seems to be more relevant to compare surfaces in the ROC space, since it yields smaller confidence regions, without sacrificing coverage.

Table 2.1: Empirical coverage probabilities for 95% empirical regions according to the bootstrap methods.

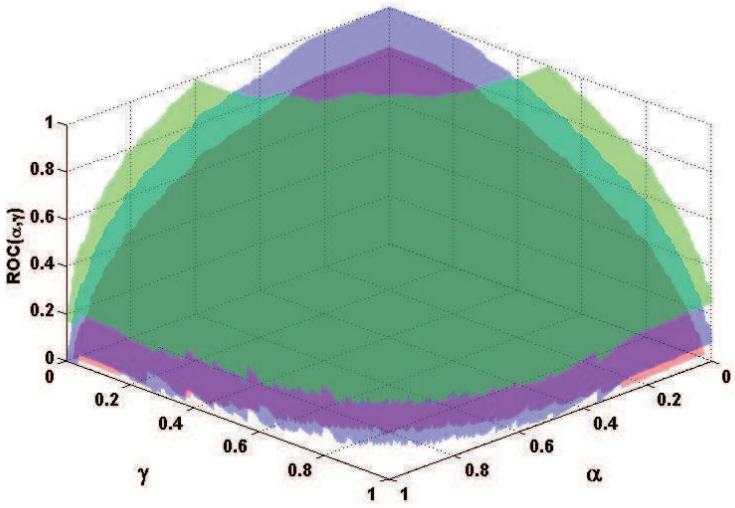
Method	δ	coverage (%)
Naive Bootstrap $\|r_n\|_\infty$	0.2147	0.9210
Smoothed Bootstrap $\|r_n\|_\infty$	0.2203	0.9450
Naive Bootstrap d_B	0.1134	0.6860
Smoothed Bootstrap d_B	0.1401	0.9510

2.3.2 Real data

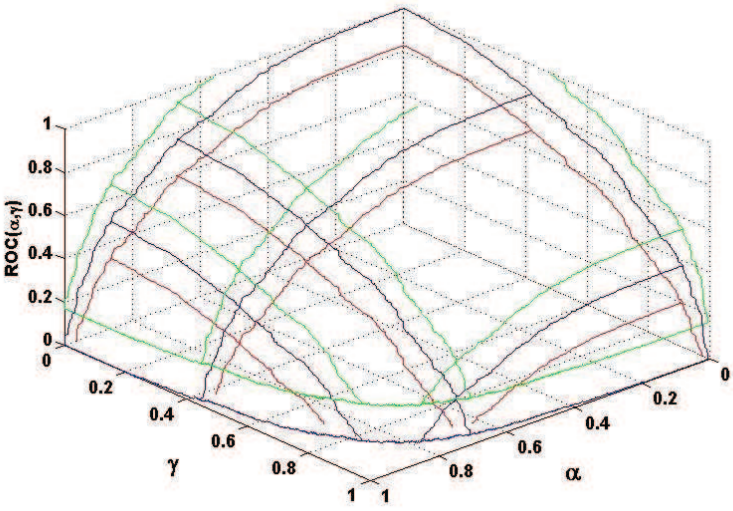
The purpose of this subsection is to compare scoring functions through a real data set, in the area of psychometrics. The "Social Workers Decisions" (SWD) data set collects real-world assessments of qualified social workers regarding the risk facing children if they stayed with their families at home. This evaluation of risk assessment is often presented to judicial courts to help decide what is in the best interest of an alleged abused or neglected child [David, 2008]. For each of the 1000 persons, 10 features are observed, together with one discrete output ranging from 1 to 3 (classes 3 and 4 of the original dataset have been merged).

The experiment has been implemented as follows, in two steps. The sample available is split into a training dataset $\mathcal{D}_e$ and a test dataset $\mathcal{D}_t$. A scoring function is first learnt using the training set $\mathcal{D}_e$ and secondly applied to the test data set to build a confidence region by means of the smoothed bootstrap algorithm described in section 2.2.2. We used two methods to build ranking functions. The first technique is linear discriminant analysis (LDA) for multi-class data, while the second one is the *the Kendall aggregation approach for multi-class ranking* developed in Chapter 4. For the LDA, we proceed the following way. After having performed LDA from data $\mathcal{D}_e$, we computed estimates of the posterior probabilities and form the scoring rule is $\hat{\eta} = \hat{\eta}_1 + 2\hat{\eta}_2 + 3\hat{\eta}_3$. The related test ROC surface is displayed in Fig. 2.3. The *Kendall aggregation approach for multi-class ranking* (KAMR) is implemented from the TreeRank procedure proposed in [Clémençon *et al.*, 2011a], based on data $\mathcal{D}_e$. This procedure is fully described in Chapter 4. The corresponding test ROC surface is plotted in Fig. 2.4. Then, we built confidence regions for the $\|\|_\infty$ and the symmetrized distance d_B using smooth bootstrap (SB) and naive bootstrap (NB). For each of these two ranking techniques, we reported the results in Table 2.2.

Looking at Fig. 2.3 and 2.4, the test ROC surface associated to the KAMR procedure is significantly better (*i.e.* higher) than the one associated to the LDA procedure. Additionally, as shown by Table 2.2, the smooth bootstrap estimate of the distance in sup norm (respectively, of the distance d_B) between the ROC surfaces of these scoring rules is equal to 0.5497 (resp., to 0.2358). These values



(a)$\|.\|_\infty$ Confidence region of the blue ROC surface.



(b) skeleton.

Figure 2.2: Plots of confidence regions.

are greater than the corresponding confidence parameters for smooth and naive bootstrap. These bootstrap statistics thus permit to assess that the true ROC surface related to the KMAR rule dominates everywhere that related to the LDA

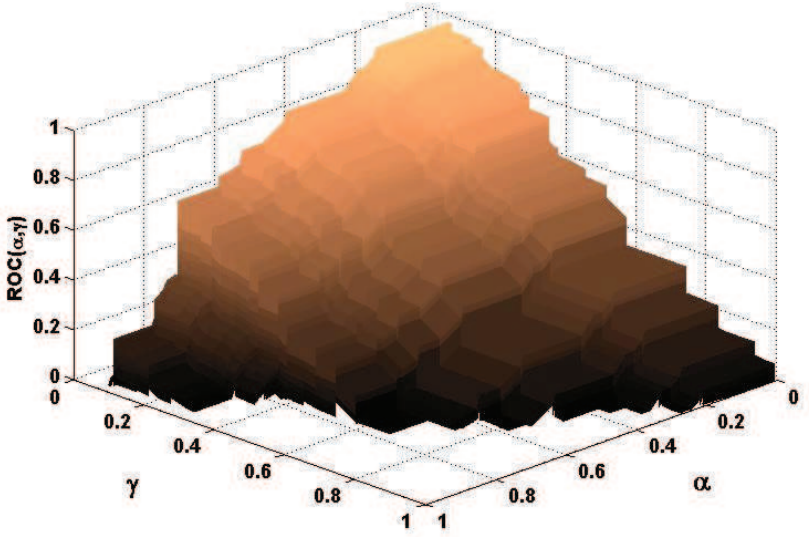


Figure 2.3: LDA ROC surface

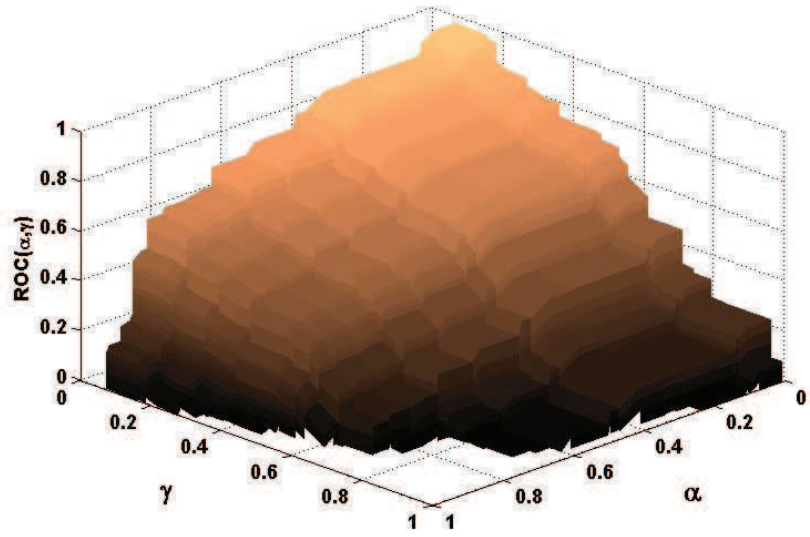


Figure 2.4: KAMR ROC surface

rule with a confidence at least 95%.

Table 2.2: Empirical size of the 95% empirical regions for the norm $\|.\|_\infty$ and d_B according to the bootstrap methods.

	LDA		KAMR	
Method	NB	SB	NB	SB
δ_∞	0.4966	0.4434	0.3949	0.3862
δ_B	0.2066	0.2144	0.1877	0.2318

2.4 Discussion

In this chapter, statistical assessment of the ROC surface, the gold standard to evaluate performance of scoring rules in ranking problems, is investigated. Asymptotic properties of a nonparametric estimator of the ROC surface of a given scoring function, based on empirical versions of the class distributions, are established. A strong approximation result for its fluctuation process is proved, with a rate of convergence of the order $O((\log\log n)^{\rho_1(\gamma)}\log^{\rho_2(\gamma)} n/\sqrt{n})$, extending the result in distribution obtained by [Li & Zhou, 2009]. We next proposed an algorithm to build confidence regions for the ROC surface of a given scoring function. Since the estimation method on which it relies involves the empirical quantile processes, it implements a smooth variant of the bootstrap technique. This approach yields a rate of convergence of the order $O_{\mathbb{P}}(n^{-2/5})$, slower than that of Gaussian approximation. Nevertheless, from a practical perspective, it permits to build confidence regions in a much more tractable way. We illustrated this through empirical experiments based on simulated/real data. They reveal in particular that the smoothing approach significantly improves on the naive bootstrap in this context.

2.5 Proofs

Proof of Theorem 2.2.1, (i) and (ii)

(i) We bound the supremum of the deviation as follows:

$$\sup_{0\leq\alpha,\gamma\leq 1} |\widehat{\mathrm{ROC}}(s,\alpha,\gamma) - \mathrm{ROC}(\mathrm{s},\alpha,\gamma)| \leq \mathrm{A}_1 + \mathrm{A}_2 + \mathrm{A}_3 + \mathrm{A}_4,$$

where

$$\begin{aligned}
A_1 &= \sup_{0\leq\alpha,\gamma\leq 1} |\hat{F}_{s,2}(\hat{F}^-_{s,3}(\gamma)) - F_{s,2}(\hat{F}^-_{s,3}(\gamma))|,\\
A_2 &= \sup_{0\leq\alpha,\gamma\leq 1} |F_{s,2}(\hat{F}^{-1}_{s,3}(\gamma)) - F_{s,2}(F^{-1}_{s,3}(\gamma))|,\\
A_3 &= \sup_{0\leq\alpha,\gamma\leq 1} |\hat{F}_{s,2}(\hat{F}^-_{1}(\alpha)) - F_{s,2}(\hat{F}^-_{1}(\alpha))|,\\
A_4 &= \sup_{0\leq\alpha,\gamma\leq 1} |F_{s,2}(\hat{F}^{-1}_{1}(\alpha)) - F_{s,2}(F^{-1}_{1}(\alpha))|
\end{aligned}$$

The terms A_1 and A_3 almost-surely vanish, by virtue of the Glivenko-Cantelli theorem. For A_2 and A_4, we use the classical inequality for the maximal deviation

in [Dvoretzky *et al.*, 1956] combined with the Borel-Cantelli theorem to obtain the desired result.

(*ii*) First, we show that $\mathbb{I}\{(\alpha,\gamma) \in \widehat{\mathcal{I}_s}\} = \mathbb{I}\{(\alpha,\gamma) \in \mathcal{I}_s\}$ a.s. when $n \to \infty$. We have $\widehat{\mathcal{I}_s} \setminus \mathcal{I}_s = \{(\alpha,\gamma) \in [0,1]^2 : \quad F_{s,3}(F_1^{-1}(\alpha)) \leq \gamma \leq \hat{F}_{s,3}(\hat{F}_{s,1}^{-1}(\alpha))\}$. Using the same argument as in (i), we obtain that $\sup_{\alpha\in[0,1]} |F_{s,3}(F_1^{-1}(\alpha) - \hat{F}_{s,3}(\hat{F}_{s,1}^{-1}(\alpha))| = 0$ a.s. when $n \to \infty$, and the desired result follows. Then, we write

$$\widehat{\mathrm{ROC}}(s,\alpha,\gamma) - \mathrm{ROC(s,}\alpha,\gamma) = \mathbb{I}\{(\alpha,\gamma) \in \mathcal{I}_\mathrm{s}\}\times$$
$$\left(\left(\hat{F}_{s,2} \circ \hat{F}_{s,3}^{-1}(1-\gamma) - \hat{F}_{s,2} \circ \hat{F}_{s,1}^{-1}(\alpha)\right) - \left(F_{s,2} \circ F_{s,3}^{-1}(1-\gamma) - F_{s,2} \circ F_{s,1}^{-1}(\alpha)\right)\right)$$
$$\text{a.s. when } n \to \infty.$$

We decompose the terms involved in the equality above as follows,

$$\left(\hat{F}_{s,2} \circ \hat{F}_{s,3}^{-1}(1-\gamma) - F_{s,2} \circ F_{s,3}^{-1}(1-\gamma)\right) - \left(\hat{F}_{s,2} \circ \hat{F}_{s,1}^{-1}(\alpha) - F_{s,2} \circ F_{s,1}^{-1}(\alpha)\right).$$

The first term can be rewriten as

$$\left(\hat{F}_{s,2} \circ \hat{F}_{s,3}^{-1}(1-\gamma) - F_{s,2} \circ \hat{F}_{s,3}^{-1}(1-\gamma) + F_{s,2} \circ \hat{F}_{s,3}^{-1}(1-\gamma) - F_{s,2} \circ F_{s,3}^{-1}(1-\gamma)\right).$$

For $\left(\hat{F}_{s,2} \circ \hat{F}_{s,3}^{-1}(1-\gamma) - F_{s,2} \circ \hat{F}_{s,3}^{-1}(1-\gamma)\right)$, we use the strong approximation result for the empirical process established in [Csorgo & Revesz, 1981],

$$\sqrt{n}(\hat{F}_{s,2}(\hat{F}_{s,3}^{-1}(1-\gamma))) - F_{s,2}(\hat{F}_{s,3}^{-1}(1-\gamma)) = \frac{\sqrt{n}}{\sqrt{n_2}} B_2^{(n)}(F_{s,2}(\hat{F}_{s,3}^{-1}(1-\gamma))) + O(log(n)/\sqrt{n}) \text{ a.s. }.$$

Applying then the LIL, we have $\frac{n}{n_2} - p_2^{-1} = O(log(n)/\sqrt{n})$. Combining the continuity of $B_2^{(n)}$ with $\sup_{0\leq\alpha,\gamma\leq1} |\hat{F}_{s,3}^{-1}(1-\gamma) - F_{s,3}^{-1}(1-\gamma)| \to 0$ a.s., we obtain

$$\sqrt{n}\left(\hat{F}_{s,2} \circ \hat{F}_{s,3}^{-1}(1-\gamma) - F_{s,2} \circ \hat{F}_3^{-1}(1-\gamma)\right) = p_2^{-1/2} B_2^{(n)}\left(F_{s,2}(F_{s,3}^{-1}(1-\gamma))\right).$$

For the difference $C_2 := \sqrt{n}(F_{s,2} \circ \hat{F}_{s,3}^{-1}(1-\gamma) - F_{s,2} \circ F_{s,3}^{-1}(1-\gamma))$ using the Taylor expansion, we get

$$C_2 = \sqrt{n} f_{s,2}(F_{s,3}^{-1}(1-\gamma))(\hat{F}_{s,3}^{-1}(1-\gamma) - F_{s,3}^{-1}(1-\gamma)) + O(1/\sqrt{n})$$

The strong approximation result for the quantile process in [Csorgo & Revesz, 1981] yields:

$$\begin{aligned}\sqrt{n}(\hat{F}_{s,3}^{-1}(1-\gamma) - F_{s,3}^{-1}(1-\gamma)) \quad &= \frac{\sqrt{n}}{\sqrt{n_3}} \frac{1}{f_{s,3}F_{s,3}^{-1}(1-\gamma)} B_3^{(n)}(1-\gamma) \\ &\quad + O((\log\log n)^{\rho_1(\xi)} \log^{\rho_2(\xi)} n/\sqrt{n}) \text{ a.s. }.\end{aligned}$$

Applying then the LIL to $n/n_2 - p_3^{-1}$, we obtain that

$$C_2 = \frac{1}{\sqrt{p_3}} \frac{f_{s,2}(F_{s,3}^{-1}(1-\gamma))}{f_{s,3}F_{s,3}^{-1}(1-\gamma)} B_3^{(n)}(1-\gamma) + O((\log\log n)^{\rho_1(\gamma)} \log^{\rho_2(\gamma)} n/\sqrt{n}) \text{ a.s. }.$$

Using the same techniques for the second term, we obtain that

$$\sqrt{n}\left(\hat{F}_{s,2}\circ\hat{F}_{s,1}^{-1}(\alpha)-F_{s,2}\circ\hat{F}_{s,3}^{-1}(\alpha)\right)=p_2^{-1}B_2^{(n)}\left(F_{s,2}(F_{s,1}^{-1}(\alpha))\right)$$

and

$$\sqrt{n}\left(F_{s,2}\circ\hat{F}_{s,1}^{-1}(1-\gamma)-F_{s,2}\circ F_{1s,}^{-1}(1-\gamma)\right)=\frac{1}{\sqrt{p_1}}\frac{f_{s,2}F_{s,1}^{-1}(\alpha))}{f_{s,1}F_{s,1}^{-1}(\alpha)}B_1^{(n)}(\alpha)$$
$$+O((\log\log n)^{\rho_1(\xi)}\log^{\rho_2(\xi)}n/\sqrt{n})\text{ a.s. }.$$

Combining these four equalities gives the desired result.

Proof of Theorem 2.2.2

Note firstly that, conditioned on $\mathcal{D}_n$ by applying Theorem 2.2.1(ii) with $\widehat{\text{ROC}}$ as a target surface (instead of ROC), one gets that $r_n^*(\alpha,\gamma)$ is, uniformly over $[0,1]^2$, almost surely equivalent to $\mathbb{I}\{\mathcal{I}_s\}z_n^*(\alpha,\gamma)$, where

$$z_n^*(\alpha,\gamma)=\left(\sqrt{\frac{n}{n_2}}B_1^{(n)}(\hat{F}_{s,2}(\hat{F}_{s,3}^{-1}(1-\gamma)))\right)+\sqrt{\frac{n}{n_3}}\frac{\tilde{f}_{s,2}(\hat{F}_{s,3}^{-1}(1-\gamma))}{\tilde{f}_{s,3}(\hat{F}_{s,3}^{-1}(1-\gamma))}B_2^{(n)}(\gamma)$$
$$-\sqrt{\frac{n}{n_2}}B_1^{(n)}(\hat{F}_{s,2}(\hat{F}_{s,1}^{-1}(\alpha)))+\sqrt{\frac{n}{n_1}}\frac{\tilde{f}_{s,2}(\hat{F}_{s,1}^{-1}(\alpha))}{\tilde{f}_{s,1}(\hat{F}_{s,1}^{-1}(\alpha))}B_3^{(n)}(\alpha)$$

with a remainder of order $O((\log\log n)^{\rho_1(\xi)}\log^{\rho_2(\xi)}n/\sqrt{n})$ almost surely. Besides,using the theorem from [Giné & Guillou, 2002], we almost surely have

$$\frac{\tilde{f}_{s,2}(\hat{F}_{s,1}^{-1}(\alpha))}{\tilde{f}_{s,1}(\hat{F}_{s,1}^{-1}(\alpha))}-\frac{f_{s,2}(\hat{F}_{s,1}^{-1}(\alpha))}{f_{s,1}(\hat{F}_{s,1}^{-1}(\alpha))}=O\left(\frac{\log(h_n^{-1})}{\sqrt{nh_n}}\right)$$

and

$$\frac{\tilde{f}_{s,2}(\hat{F}_{s,3}^{-1}(1-\gamma))}{\tilde{f}_{s,3}(\hat{F}_{s,3}^{-1}(1-\gamma))}-\frac{f_{s,2}(\hat{F}_{s,3}^{-1}(1-\gamma))}{f_{s,3}(\hat{F}_{s,3}^{-1}(1-\gamma))}=O\left(\frac{\log(h_n^{-1})}{\sqrt{nh_n}}\right)$$

under the stipulated conditions. Furthermore, from standard result on the modulus of continuity ([Shorack & Wellner, 1986]), we almost surely have

$$\sup_{\alpha\in[0,1]}|B_1^{(n)}(\hat{F}_{s,2}(\hat{F}_{s,1}^{-1}(\alpha)))-B_1^{(n)}(F_{s,2}(F_{s,1}^{-1}(\alpha)))|=O\left(\frac{\log(n)}{\sqrt{n}}\right),$$

and

$$\sup_{\gamma\in[0,1]}|B_1^{(n)}(\hat{F}_{s,2}(\hat{F}_{s,3}^{-1}(1-\gamma)))-B_1^{(n)}(F_{s,2}(F_{s,3}^{-1}(1-\gamma)))|=O\left(\frac{\log(n)}{\sqrt{n}}\right).$$

Applying the LIL to $\frac{n_1}{n}$, $\frac{n_2}{n}$ and $\frac{n_3}{n}$, it follows that almost surely, uniformly in α and γ,

$$z_n^*(\alpha,\gamma) = \left(\sqrt{\frac{1}{p_2}}B_1^{(n)}(F_{s,2}(F_{s,3}^{-1}(1-\gamma)))\right) + \sqrt{\frac{1}{p_3}}\frac{f_{s,2}(F_{s,3}^{-1}(1-\gamma))}{f_{s,3}(F_{s,3}^{-1}(1-\gamma))}B_2^{(n)}(\gamma)$$
$$- \sqrt{\frac{1}{p_2}}B_1^{(n)}(F_{s,2}(F_{s,1}^{-1}(\alpha))) + \sqrt{\frac{1}{p_1}}\frac{f_{s,2}(F_{s,1}^{-1}(\alpha))}{f_{s,1}(F_{s,1}^{-1}(\alpha))}B_3^{(n)}(\alpha) + O\left(\frac{\log(h_n^{-1})}{\sqrt{nh_n}}\right).$$

Since this result holds uniformly in α and γ, by continuous mapping theorem applied to the function $\sup_{(\alpha,\gamma)\in[0,1]^2}(.)$, we also get the results in term of distribution up to the given almost sure order. Thus, we obtain the first part of the theorem.

Proof of Theorem 2.2.3

Recall that the distance $\sqrt{n}d_B(\mathrm{ROC}, \widehat{\mathrm{ROC}}, \alpha, \gamma)$ involves the minimum over 5 quantities. We treat each of the latter separately. The first one is $\sqrt{n}(\mathrm{ROC}(\alpha,\gamma) - \widehat{\mathrm{ROC}}(\alpha,\gamma))$ and its strong approximation is given in Theorem 2. Additionally, in Theorem 3, it is shown that the related bootstrap process, *i.e.* $\sqrt{n}(\widetilde{\mathrm{ROC}}(\alpha,\gamma) - \widehat{\mathrm{ROC}}(\alpha,\gamma))$ has the same law, up to the additional term $O\left(\frac{\log(h_n^{-1})}{\sqrt{nh_n}}\right)$. For the four other terms, we have to establish strong approximations and this boils down to mimic the proofs of Theorems 2 and 3 for each term. Notice that the generalized inverse of the cdf $F_{s,2}$ is involved in these terms, which explains why assumption $\mathbf{A_3}$ is required. Once each term is controlled, the continuous mapping theorem yields the desired result.

Chapter 3

Subsampling the VUS criterion and empirical maximization

In statistical learning theory, the paradigmatic approach to predictive problems is to use data-based estimates of the prediction error to select a decision rule from a class of candidates. In classification/regression, such estimates are sample mean statistics and the theory of *Empirical Risk Minimization* (ERM in abbreviated form) has been originally developed in this situation, relying essentially on the study of maximal deviations between these empirical averages and their expectations. The tools used for this purpose are mainly concentration inequalities for empirical processes; see [Ledoux & Talagrand, 1991] for instance. One may refer to [Boucheron *et al.*, 2005] for a recent account of the theory of classification.

Recently, a variety of learning issues, where natural empirical risk estimates are no longer basic sample mean statistics, have received a good deal of attention in the machine-learning literature, requiring to extend the ERM approach. Indeed, in certain problems such as *supervised ranking* [Clémençon *et al.*, 2008], *learning on graphs* [Biau & Bleakley, 2006] or *pairwise dissimilarity-based clustering* [Clémençon, 2011], statistical counterparts of the risk are of the form of (generalized) U-statistics; see [Lee, 1990]. Such empirical functionals are computed by averaging over tuples of sampling observations, exhibiting thus a complex dependence structure. *Linearization techniques* (see [Hoeffding, 1948]) are the main ingredient in investigating the behavior of empirical risk minimizers in this setting, the latter permitting to establish probabilistic upper bounds for the maximal deviation of collection of centered U-statistics under adequate conditions by reducing the study to that of standard empirical processes.

However, while the ERM theory based on minimization of U-statistics is now consolidated, putting this approach in practice generally leads to face significant computational difficulties, not sufficiently well documented in the machine-learning literature. In many concrete cases, the mere computation of the risk involves a summation which extends over an extremely high number of tuples and runs out of time or memory on most machines. It is the major purpose of this chapter to study how a simplistic sampling technique (*i.e.* drawing with replacement) applied to risk estimation, as originally proposed by [Blom, 1976] in the context of asymptotic pointwise estimation, may efficiently remedy this issue without damaging too much the "reduced variance" property of the estimates, while preserving the learning rates (including "fast-rate" situations). Applications to *supervised ranking* is considered

here in order to illustrate this remarkable phenomenon.

The chapter is structured as follows. In section 3.1, we explain the interest of empirical maximization of U-statistics in the case of multipartite ranking. In chapter 3.2, we present the resampling procedure and we state the main theorem of this chapter, a concentration inequalities for U-processes. In section 3.3, we exploit this theorem in the framework of ranking and we give numerical illustrations of the resampling procedure. Mathematical proofs are postponed to section 3.5.

3.1 Motivation

As we have seen in chapter 1, the main criterion in multipartite ranking is the VUS (see 1.2.5) and its empirical counterpart takes the form of a U-statistic. When the scoring function has no ties (almost surely), if K independent samples, of independent copies of the r.v. $X^{(k)}$ respectively, are available,

$$X_1^{(k)} \ \ldots, \ X_{n_k}^{(k)} \text{ with } n_k \geq 1 \text{ for } 1 \leq k \leq K, \tag{3.1}$$

the empirical VUS can be written as :

$$\widehat{\mathrm{VUS}}_{\mathbf{n}}(s) = \frac{\sum_{i_1=1}^{n_1} \cdots \sum_{i_K=1}^{n_K} \mathbb{I}\left\{s(X_{i_1}^{(1)}) < \ldots < s(X_{i_K}^{(K)})\right\}}{n_1 \times \cdots \times n_K}, \tag{3.2}$$

where $\mathbf{n} = (n_1, \ \ldots, \ n_K)$. When the scoring function has ties, the kernel of the U-statistic has 2^{K-1} terms, see remark 1.2.2. The performance of empirical maximizers of the quantity (3.2) (or of variants of the latter performance measure) over a class $\mathcal{S}$ of scoring function candidates has been investigated in several papers, mainly in the *bipartite* context (*i.e.* for $K = 2$), under various complexity assumptions for $\mathcal{S}$; see [Agarwal *et al.*, 2005, Clémençon *et al.*, 2008] among others. Although the computation of the VUS can be done in $O(n \ln n)$ using the algorithm in Annex 1.5, in practice algorithms are based on finding $\hat{s}$ that maximizes

$$\arg\max_{s \in \mathcal{S}_0} \widehat{\mathrm{VUS}}_{\mathbf{n}}(s)$$

where $\mathcal{S}_0$ is a collection of scoring function. Solving such an optimization problem is of complexity $n_1 \times \cdots \times n_K$.

In a variety of applications (information retrieval, design of recommender systems for instance), the number of classes K and/or the sample sizes n_k are fairly large, so that the complexity $n_1 \times \cdots \times n_K$ is prohibitive. As an illustration, one may refer to the public databases LETOR (available at `http://research.microsoft.com/~letor/`), which can be used to evaluate search engines for ranking documents according to their degree of pertinence for specific requests in particular (see [Liu *et al.*, 2007]), where $K = 5$ and the sample sizes are very huge for most queries. Datasets released for recent competitions, such as the

Yahoo! Labs "Learning to Rank" challenge in 2010 or the *KDD Cup Orange* challenge in 2009, provide other examples of such situations. In the *KDD Cup Orange* challenge, where submissions were evaluated based on the AUC performance, the computation of the empirical version of the criterion required to average over 10^{12} pairs approximately, making "pairwise classification" approaches inatractable (unless the sampling technique promoted here and analyzed in the subsequent section is used).

3.2 Uniform approximation of generalized U-statistics through sampling

As will be seen below, the statistics considered in the previous section are (generalized) U-statistics, which can be *uniformly* approximated by Monte-Carlo versions whose computation cost is drastically reduced. This will be next proved to be an essential tool for investigating the performance of decision rules learnt through optimization of such empirical quantities.

3.2.1 Definitions and key properties

For the sake clarity, we recall the definition of generalized U-statistics, the simplest extensions of standard sample mean statistics. Properties and asymptotic theory of U-statistics can be found in [Lee, 1990].

Definition 3.2.1. (GENERALIZED U-STATISTIC) *Let $K \geq 1$ and $(d_1, \ldots, d_K) \in \mathbb{N}^{*K}$. Let $(X_1^{(k)}, \ldots, X_{n_k}^{(k)})$, $1 \leq k \leq K$, be K independent samples of i.i.d. random variables, taking their values in some space $\mathcal{X}_k$ with distribution $F_k(dx)$ respectively. The generalized (or K-sample) U-statistic of degrees $(d_1, \ldots, d_K)$ with kernel $H : \mathcal{X}_1^{d_1} \times \cdots \times \mathcal{X}_K^{d_K} \to \mathbb{R}$, square integrable with respect to the probability distribution $\mu = F_1^{\otimes d_1} \otimes \cdots \otimes F_K^{\otimes d_K}$, is defined as*

$$U_{\mathbf{n}}(H) = \frac{\sum_{I_1} \cdots \sum_{I_K} H(\mathbf{X}_{I_1}^{(1)}; \mathbf{X}_{I_2}^{(2)}; \ldots; \mathbf{X}_{I_K}^{(K)})}{\binom{n_1}{d_1} \times \cdots \times \binom{n_K}{d_K}}, \tag{3.3}$$

where the symbol $\sum_{I_k}$ refers to summation over all $\binom{n_k}{d_k}$ subsets $\mathbf{X}_{I_k}^{(k)} = (X_{i_1}^{(k)}, \ldots, X_{i_{d_k}}^{(k)})$ related to a set I_k of d_k indices $1 \leq i_1 < \ldots < i_{d_k} \leq n_k$. It is said symmetric when H is permutation symmetric in each set of d_k arguments $\mathbf{X}_{I_k}^{(k)}$.

Coming back to the example of the previous section, we observe that, for a fixed scoring function $s(x)$, the quantity (3.2) is a K-sample U-statistic of degree $(1, 1, \ldots, 1)$ with kernel given by:

$$\mathrm{H}_s(x_1, \ldots, x_K) = \mathbb{I}_{\{s(x_1) < s(x_2) < \cdots < s(x_K)\}}$$

for $(x_1, \ldots, x_K) \in \mathcal{X}^K$.

Beyond this example, many statistics used for pointwise estimation or hypothesis testing are actually U-statistics (*e.g.* the sample variance, the Gini mean difference, the Wilcoxon Mann-Whitney statistic, Kendall tau), their popularity mainly arise from their "reduced variance" property: the statistic $U_{\mathbf{n}}(H)$ has minimum variance among all unbiased estimators of the parameter

$$\theta(H) = \mathbb{E}[H(X_1^{(1)}, \ldots, X_{d_1}^{(1)}, \ldots, X_1^{(K)}, \ldots, X_{d_k}^{(K)})].$$

Asymptotic.

Classically, the limit properties of these statistics (LLN, CLT, *etc.*) are investigated in an asymptotic framework stipulating that, as the full sample size

$$n \overset{def}{=} n_1 + \ldots + n_K$$

tends to infinity, we have: $n_k/n \to \lambda_k > 0$ for $k = 1, \ldots, K$. They can be established by means of a linearization technique (see [Hoeffding, 1948]), permitting to write $U_{\mathbf{n}}(H)$ as a sum of K basic sample mean statistics (of the order $O_{\mathbb{P}}(1/\sqrt{n})$ each, after recentering), plus possible degenerate terms (termed *degenerate* U-statistics). This method is extensively used in [Clémençon *et al.*, 2008] for instance.

As previously seen on the running example considered in this chapter, in practice, the number $\prod_{k=1}^{K} \binom{n_k}{d_k}$ of terms to be summed up to compute (3.3) is generally prohibitive. As a remedy to this computational issue, in the seminal contribution [Blom, 1976], the concept of *incomplete generalized U-statistic* has been introduced, where the summation in formula (3.3) is replaced by a summation involving much less terms, extending over low cardinality subsets of the $\binom{n_k}{d_k}$ d_k-tuples of indices, $1 \leq k \leq K$, solely. In the simplest formulation, the subsets of indices are obtained by sampling with replacement, leading to the following definition.

Definition 3.2.2. (Incomplete Generalized U-statistic) *Let $B \geq 1$. The incomplete version of the U-statistic (3.3) based on B terms is defined by:*

$$\widetilde{U}_B(H) = \frac{1}{B} \sum_{(I_1, \ldots, I_K) \in \mathcal{D}_B} H(X_{I_1}^{(1)}, \ldots, X_{I_K}^{(K)}), \tag{3.4}$$

where $\mathcal{D}_B$ is a set of cardinality B built by sampling with replacement in the set $\Lambda = \{((i_1^{(1)}, \ldots, i_{d_1}^{(1)}), \ldots, (i_1^{(K)}, \ldots, i_{d_K}^{(K)})) : 1 \leq i_1^{(k)} < \ldots < i_{d_k}^{(k)} \leq n_k, 1 \leq k \leq K\}$.

Remark 3.2.1. (Alternative sampling schemes.) *We point out that, as proposed in [Janson, 1984], other sampling schemes could be considered, sampling without replacement or Bernoulli sampling in particular. The results of this chapter could be extended to these situations.*

In practice, B should be chosen much smaller than the cardinality of Λ, namely $\#\Lambda = \prod_{k=1}^{K} \binom{n_k}{d_k}$, in order to overcome the computational issue previously mentioned. We emphasize that the cost related to the computation of the value taken by the kernel H at a given point $(x_{I_1}^{(1)}, \ldots, x_{I_K}^{(K)})$ depending on the form of H is not considered here, focus is on the number of terms involved in the summation solely. As an estimator of $\theta(H)$, the statistic (3.4) is still unbiased but its variance is naturally larger than that of (3.3). Precisely, we have

$$\mathrm{Var}(\widetilde{U}_B(H)) = (1 - 1/B)\mathrm{Var}(U_{\mathbf{n}}(H)) + O(1/B),$$

as $B \to +\infty$; refer to [Lee, 1990] (see p. 193 therein). Incidentally, we underline that the empirical variance of (3.3) is not easy to compute neither since it involves summing approximately $\#\Lambda$ terms and bootstrap techniques should be used for this purpose, as proposed in [Bertail & Tressou, 2006]. The asymptotic properties of incomplete U-statistics have been investigated in several articles; see [Brown & Kildea, 1978, Enqvist, 1978, Janson, 1984]. The angle embraced in the present chapter is of quite different nature, the key idea we promote here is to use incomplete versions of collections of U-statistics in learning problems such as those described in section 3.1. The result established in the next section shows that this approach solves the numerical problem, while not damaging the learning rates.

3.2.2 Main result - Uniform approximation of U-statistics by incomplete U-statistics

Under some assumptions on the collection $\mathcal{H}$ of (symmetric) kernels H considered, concentration results established for U-processes (*i.e.* collections of U-statistics) may extend to their incomplete versions, as revealed by the following theorem. We consider the (not that restrictive) situation where the class $\mathcal{H}$ of kernels is a VC major class of functions of finite Vapnik-Chervonenkis dimension; see [Dudley, 1999].

Theorem 3.2.1. (Maximal deviation) *Let $\mathcal{H}$ be a collection of bounded symmetric kernels on $\Omega = \prod_{k=1}^{K} \mathcal{X}_k^{d_k}$ of finite* VC *dimension $\mathcal{V} < +\infty$. We set $\mathcal{M}_{\mathcal{H}} = \sup_{(H,x)\in\mathcal{H}\times\mathcal{X}} |H(x)|$. Then, the following assertions hold.*

(i) For all $\eta > 0$, we have: $\forall \mathbf{n} = (n_1, \ldots, n_K) \in \mathbb{N}^{*K}$, $\forall B \geq 1$,

$$\mathbb{P}\left\{ \sup_{H\in\mathcal{H}} \left| \widetilde{U}_B(H) - U_{\mathbf{n}}(H) \right| > \eta \right\} \leq 2(1 + \#\Lambda)^{V} \times e^{-B\eta^2/\mathcal{M}_{\mathcal{H}}^2}.$$

(ii) For all $\delta \in (0,1)$, with probability at least $1-\delta$, we have: $\forall n_k \geq 1$, $1 \leq k \leq K$,

$$\frac{1}{\mathcal{M}_{\mathcal{H}}} \sup_{H\in\mathcal{H}} \left| \widetilde{U}_B(H) - \mathbb{E}\left[\widetilde{U}_B(H)\right] \right| \leq 2\sqrt{\frac{2\mathcal{V}\log(1+\kappa)}{\kappa}} + \sqrt{\frac{\log(2/\delta)}{\kappa}} + \sqrt{\frac{\mathcal{V}\log(1+\#\Lambda) + \log(4/\delta)}{B}}, \quad (3.5)$$

where $\kappa = \min\{\lfloor n_1/d_1 \rfloor, \ldots, \lfloor n_K/d_K \rfloor\}$ *and* $\lfloor x \rfloor$ *denotes the integer part of any real number* x.

Refer to the Appendix for the proof. The bounds stated above show that, for a number $B = B_n$ of terms tending to infinity as $n \to +\infty$ at a rate $O(n)$, the maximal deviation $\sup_{H \in \mathcal{H}} |\widetilde{U}_B(H) - \theta(H)|$ is asymptotically of the order $O_{\mathbb{P}}(n^{-1/2})$, just like $\sup_{H \in \mathcal{H}} |U_{\mathbf{n}}(H) - \theta(H)|$. Remarkably, except in the case $K = 1$ and $d_K = 1$ solely, using such incomplete U-statistics thus yields a significant gain in terms of computational cost and preserves the order of the probabilistic upper bounds for the uniform deviation.

3.3 Maximization of the VUS

We now discuss the consequences of Theorem 3.2.1 through the problem of maximizing the VUS introduced in section 3.1 (notice that, in this case, we have $\mathcal{M}_{\mathcal{H}} = 1$). Beyond theoretical guarantees, the performance of algorithms based on incomplete versions of the empirical counterpart of the functional of interest is illustrated by numerical results, supporting the efficiency of the sampling approach promoted in this chapter in the machine-learning context.

3.3.1 Sampling the risk in K-partite ranking

We come back to the ranking framework. Here, the full replacement size is $n = n_1 + \ldots + n_K$. Let $(b_1, b_2, \ldots, b_K)$ be a sequence of nonnegative integers such that:

$$\forall k \in \{1, \ldots, K\}, \ \ b_k \sim n_k^{1/K} \sim n^{1/K} \text{ as } n \to +\infty.$$

The sampling scheme consists, for $1 \leq k \leq K$, of drawing with replacement b_k observations in the sample k: $X_{i_1}^{(k)}, \ldots, X_{i_{b_k}}^{(k)}$. Set $B = b_1 \times \cdots \times b_K$. Based on the sampled data, we compute the following estimate of the VUS criterion

$$\widetilde{\mathrm{VUS}}_B(s) = \frac{1}{B} \sum_{l_1=1}^{b_1} \cdots \sum_{l_K=1}^{b_K} \mathbb{I}\left\{ s(X_{i_{l_1}^{(1)}}^{(1)}) < \ldots < s(X_{i_{l_K}^{(K)}}^{(K)}) \right\},$$

and consider the maximizer over a class $\mathcal{S}$ of scoring function candidates:

$$\widehat{s}_B = \arg\max_{s \in \mathcal{S}} \widetilde{\mathrm{VUS}}_B(s). \tag{3.6}$$

The following result provides a rate bound for the VUS of the scoring function above (neglecting the bias term).

Corollary 3.3.1. *Suppose that* $\mathcal{S}$ *is a* VC *major class of functions of finite* VC *dimension* $\mathcal{V} < +\infty$. *Then, for all* $\delta \in (0, 1)$, *we have with probability at least* $1 - \delta$: $\forall \mathbf{n} \in \mathbb{N}^{*K}$,

$$\max_{s \in \mathcal{S}} \mathrm{VUS(s)} - \mathrm{VUS}(\widehat{\mathrm{s}}_{\mathrm{B}}) \leq \mathrm{c}\sqrt{\frac{\mathcal{V}\log(\#\Lambda/\delta)}{\mathrm{n}}}, \tag{3.7}$$

for some constant $c < +\infty$.

The proof immediately derives from Theorem 3.2.1, details are left to the reader. One should pay attention to the fact that the deficit of VUS of the rule obtained through maximization of statistics computed by averaging $O(n)$ terms is thus of the same order as that of $\arg\max_{s\in\mathcal{S}} \widehat{\mathrm{VUS}}_{\mathbf{n}}(s)$, whose computation requires to evaluate averages extending over $O(n^K)$ terms.

Remark 3.3.1. (On fast rates) *In the bipartite setup (i.e.* $K = 2$*), situations where fast rates of convergence can be achieved by* $\arg\max_{s\in\mathcal{S}} \widehat{\mathrm{VUS}}_{\mathbf{n}}(s)$ *have been exhibited; see [Clémençon et al., 2008]. We point out that, in these situations, the same rate bounds can be attained by* $\widehat{s}_B$*, at the price of a higher computational cost (i.e. of a larger asymptotic order for* B*) however.*

3.3.2 Illustrations

Here, we illustrate the methodology from a practical point of view through a simulated dataset and a real dataset.

A numerical example with $K = 5$.

As an illustration, we display below some results related to the performance of the algorithm SVMrank (implemented with default parameters, linear kernel and $C = 20$; see [Herbrich *et al.*, 2000] and Chapter 6 for more details) using the SVMlight implementation available at `http://svmlight.joachims.org/`. We simulated a mixture of 5 Gaussian distributions on $\mathbb{R}^2$ with means $m_1, \ldots, m_5$ respectively, where $m_i = (i/6,\ i/6)$ for $1 \leq i \leq 5$, and same covariance matrix $(1/15,\ 0;\ 0,\ 1/15)$, so that an optimal scoring function (w.r.t. the VUS criterion) is given by: $s(x, y) = x + y$ for all $(x, y) \in \mathbb{R}^2$. We independently drew 50 training samples of size $n = 10\,000$ (2 000 per class) and a test sample of size 10 000. Inside each class, we drew with replacement b observations and formed the dataset $\mathcal{D}_b$, for $b = 20,\ 100$. The results, averaged over the 50 replications, are reported in Table 3.1.

Table 3.1: Comparison of the empirical VUS : $\mathrm{VUS}^* = 0.1525$

% of data	1%	5%	100%
$\overline{L}$	0.1497	0.1520	0.1524
$\widehat{\sigma}$	0.0041	0.0008	0.0002
time (in seconds)	10	200	148523

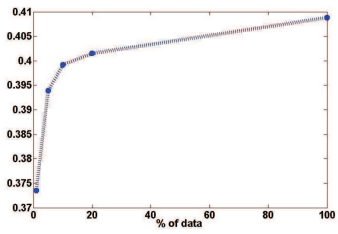
We see that, even for $b = 20$ (*i.e* 1% of the data), the performance is close to the optimum VUS^* for a computation time reduced by a factor 10000. For $b = 100$ (*i.e* 5% of the data), it is quasi-optimal, with a gain in time of a factor greater than 500.

LETOR4.0 datasets.

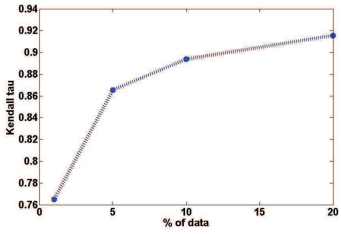
We also implemented the approach promoted in this chapter on the benchmark LETOR datasets, (see `research.microsoft.com/en-us/um/people/letor/`), by means of the same ranking algorithm as that used in the previous experiment. To be more precise, we used the two query sets MQ2007 and MQ2008, where pairs "page-query" assigned to a discrete label ranging from 0 to 2 (*i.e.* "non-relevant" - "relevant" - "extremely relevant") are gathered. In both datasets, 46 features are collected, over 69 623 instances in MQ2007 and over 15 211 instances in MQ2008. In each case, an estimate of the ranking risk L has been computed through 5 replications of a five-fold cross validation procedure, the results (mean and standard error) are reported in Tables 3.2 and 3.3. We also compute the Kendall τ statistic $\widehat{\tau}$ between the resulting rankings (recall that it ranges from -1 "full disagreement" to $+1$ "full agreement"), when using 1% ,5%, 10%, 20% and 100% of the data in each of the $K = 3$ samples. The results are reported in the Tables 3.2 and 3.3.

Table 3.2: Empirical VUS : "LETOR 2008".

%	1%	5%	10%	20%	100%
$\overline{L}$	0.3735	0.3939	0.3992	0.4015	0.4088
$\widehat{\sigma}$	0.0038	0.0040	0.0025	0.0027	0.0006
$\widehat{\tau}$	0.7648	0.8653	0.8937	0.9154	1



(a) Empirical VUS for SVMRANK based on 1%, 5%, 10%, 20% and 100% of the "LETOR 2008" dataset.



(b) Empirical Kendall τ between the scoring functions learned with 1% ,5%, 10% and 20% and the function learned using all the training set.

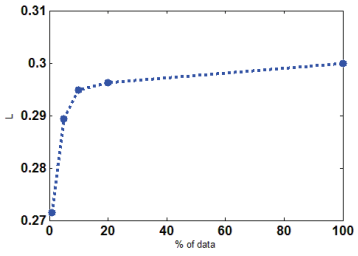
In both experiments, we observe that, as b_k/n_k increase, the ranking performance of the rules produced by the algorithm gets rapidly closer and closer to that of the ranking rule based on the whole dataset.

3.4 Conclusion

Though of great simplicity, the results stated in this chapter are of crucial importance in practice in the "big data" era. They hopefully shed light on tractable strategies for implementing learning techniques, when the (risk) functional has a

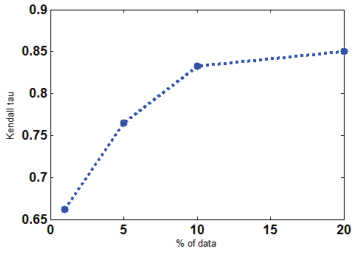
Table 3.3: Empirical VUS : "LETOR 2007".

%	1%	5%	10%	20%	100%
$\overline{L}$	0.2715	0.2894	0.2949	0.2963	0.3000
$\widehat{\sigma}$	0.0077	0.0027	0.0017	0.0019	0.0004
$\widehat{\tau}$	0.6621	0.7651	0.8328	0.8501	1



(a)

Empirical VUS for SVMRANK based on 1%, 5%, 10%, 20% and 100% of the "LETOR 2008" dataset.



(b)

Empirical Kendall τ between the scoring functions learned with 1% ,5%, 10% and 20% and the function learned using all the training set.

statistical counterpart which is of the form of a U-statistic. Whereas the theoretical properties of decision rules based on optimizing such statistics are becoming well-documented in the machine-learning literature, computational issues related to the practical implementation of learning algorithms dedicated to these optimization problems had not been tackled, to the best of our knowledge. The essential contribution of this chapter is to provide theoretical/empirical evidence that using *incomplete U-statistics* as estimates of the criterion of interest may provide a simple and elegant way of dramatically reducing computational cost in practice, while yielding nearly optimal solutions. The analysis, carried out here in a finite VC dimension framework, suggests to investigate next the use of such statistics for model selection issues and to study concentration properties of *weighted multinomial random variables* involved in the maximal deviation between U-statistics and their incomplete versions. Moreover, the analysis is directly applicable to the multipartite ranking problem and lead to an estimation of the VUS for scoring function so this methodology can be used to compare performance of scoring functions over very huge dataset.

3.5 Proofs

Appendix - Proof of Theorem 3.2.1

For convenience, we introduce the random sequence $\varepsilon = ((\varepsilon_k(I))_{I\in\Lambda})_{1\leq k\leq B}$, where $\varepsilon_k(I)$ is equal to 1 if the tuple $I = (I_1, \ldots, I_K)$ has been selected at the k-th

draw and to 0 otherwise: the ε_k's are i.i.d. random vectors and, for all $(k, I) \in \{1, \ldots, B\} \times \Lambda$, the r.v. $\varepsilon_k(I)$ has a Bernoulli distribution with parameter $1/\#\Lambda$. We also set $\mathbf{X}_I = (\mathbf{X}_{I_1}^{(1)}, \ldots, \mathbf{X}_{I_K}^{(K)})$ for any I in Λ. Equipped with these notations, observe first that one may write: $\forall B \geq 1$, $\forall \mathbf{n} \in \mathbb{N}^{*K}$,

$$\widetilde{U}_B(H) - U_{\mathbf{n}}(H) = \frac{1}{B} \sum_{k=1}^{B} \mathcal{Z}_k(H),$$

where $\mathcal{Z}_k(H) = \sum_{I \in \Lambda} (\varepsilon_k(I) - 1/\#\Lambda) H(\mathbf{X}_I)$ for any $(k, I) \in \{1, \ldots, B\} \times \Lambda$. It follows from the independence between the $\mathbf{X}_I$'s and the $\varepsilon(I)$'s that, for all $H \in \mathcal{H}$, conditioned upon the $\mathbf{X}_I$'s, the variables $\mathcal{Z}_1(H), \ldots, \mathcal{Z}_B(H)$ are independent, centered and almost-surely bounded by $2\mathcal{M}_{\mathcal{H}}$ (notice that $\sum_{I \in \Lambda} \varepsilon_k(I) = 1$ for all $k \geq 1$). By virtue of Sauer's lemma, since $\mathcal{H}$ is a VC major class with finite VC dimension $\mathcal{V}$, we have, for fixed $\mathbf{X}_I$'s:

$$\#\{(H(\mathbf{X}_I))_{I \in \Lambda} : \ H \in \mathcal{H}\} \leq (1 + \#\Lambda)^V.$$

Hence, conditioned upon the $\mathbf{X}_I$'s, using the union bound and next Hoeffding's inequality applied to the independent sequence $\mathcal{Z}_1(H), \ldots, \mathcal{Z}_B(H)$, for all $\eta > 0$, we obtain that:

$$\begin{aligned}
\mathbb{P}\left\{ \sup_{H \in \mathcal{H}} \left| \widetilde{U}_B(H) - U_{\mathbf{n}}(H) \right| > \eta \ \middle| \ (\mathbf{X}_I)_{I \in \Lambda} \right\} & \\
\leq \mathbb{P}\left\{ \sup_{H \in \mathcal{H}} \left| \frac{1}{B} \sum_{k=1}^{B} \mathcal{Z}_k(H) \right| > \eta \mid (\mathbf{X}_I)_{I \in \Lambda} \right\} & \\
\leq 2(1 + \#\Lambda)^V e^{-B\eta^2/\mathcal{M}_{\mathcal{H}}^2}, &
\end{aligned}$$

which proves the first assertion of the theorem. Notice that this can be formulated: for any $\delta \in (0, 1)$, we have with probability at least $1 - \delta$:

$$\sup_{H \in \mathcal{H}} \left| \widetilde{U}_B(H) - U_{\mathbf{n}}(H) \right| \leq \mathcal{M}_{\mathcal{H}} \times \sqrt{\frac{V \log(1 + \#\Lambda) + \log(2/\delta)}{B}}.$$

The second part of the theorem straightforwardly results from the first part combined with the following result, which extends Corollary 3 in [Clémençon *et al.*, 2008] to the K-sample situation.

Lemma 3.5.1. *Suppose that Theorem 3.2.1's hypotheses are fulfilled. For all $\delta \in (0, 1)$, we have with probability at least $1 - \delta$,*

$$\frac{1}{\mathcal{M}_{\mathcal{H}}} \sup_{H \in \mathcal{H}} |U_{\mathbf{n}}(H) - \theta(H)| \leq 2\sqrt{\frac{2V \log(1 + \kappa)}{\kappa}} + \sqrt{\frac{\log(1/\delta)}{\kappa}}.$$

Proof. Set $\kappa = \min\{\lfloor n_1/d_1 \rfloor, \ldots, \lfloor n_K/d_K \rfloor\}$ and let

$$\begin{aligned}\kappa^{-1} V_H \left(X_1^{(1)}, \ldots, X_{n_1}^{(1)}, \ldots, X_1^{(K)}, \ldots, X_{n_K}^{(K)} \right) = \\ H \left(X_1^{(1)}, \ldots, X_{d_1}^{(1)}, \ldots, X_1^{(K)}, \ldots, X_{d_K}^{(K)} \right) \\ + H \left(X_{d_1+1}^{(1)}, \ldots, X_{2d_1}^{(1)}, \ldots, X_{d_K+1}^{(K)}, \ldots, X_{2d_K}^{(K)} \right) + \ldots \\ + H \left(X_{\kappa d_1 - d_1 + 1}^{(1)}, \ldots, X_{\kappa d_K - d_K + 1}^{(K)}, \ldots, X_{\kappa d_K}^{(K)} \right),\end{aligned}$$

for any $H \in \mathcal{H}$. Recall that the K-sample U-statistic $U_{\mathbf{n}}(H)$ can be expressed as

$$U_{\mathbf{n}}(H) = \frac{1}{n_1! \cdots n_K!} \times \sum_{\sigma_1 \in \mathfrak{S}_{n_1}, \ldots, \sigma_K \in \mathfrak{S}_{n_K}} V \left(X_{\sigma_1(1)}^{(1)}, \ldots, X_{\sigma_K(n_K)}^{(K)} \right),$$

where $\mathfrak{S}_m$ denotes the symmetric group of order m for any $m \geq 1$. This representation as an average of sums of κ independent terms is known as the (first) Hoeffding's decomposition; see [Hoeffding, 1948]. Then, using Jensen's inequality in particular, one may easily show that, for any nondecreasing convex function $\psi : \mathbb{R}_+ \to \mathbb{R}$, the quantity $\mathbb{E}[\psi(\sup_{H \in \mathcal{H}} |U_{\mathbf{n}}(\bar{H})|)]$ is bounded by

$$\mathbb{E} \left[\psi \left(\sup_{H \in \mathcal{H}} \left| V_{\bar{H}}(X_1^{(1)}, \ldots, X_{n_K}^{(K)}) \right| \right) \right],$$

where we set $\bar{H} = H - \theta(H)$ for all $H \in \mathcal{H}$. Now, using standard symmetrization and randomization arguments (see [Giné & Zinn, 1984] for instance) and the bound above, we obtain that

$$\mathbb{E} \left[\psi \left(\sup_{H \in \mathcal{H}} \left| U_{\mathbf{n}}(\bar{H}) \right| \right) \right] \leq \mathbb{E} \left[\psi \left(2\mathcal{R}_\kappa \right) \right], \tag{3.8}$$

where

$$\mathcal{R}_\kappa = \sup_{H \in \mathcal{H}} \frac{1}{\kappa} \sum_{l=1}^{\kappa} \varepsilon_l H \left(X_{(l-1)d_1+1}^{(1)}, \ldots, X_{ld_K}^{(K)} \right),$$

is a Rademacher average based on the Rademacher chaos $\varepsilon_1, \ldots, \varepsilon_\kappa$ (independent random symmetric sign variables), independent from the $X_i^{(k)}$'s. We now apply the bounded difference inequality (see [McDiarmid, 1989]) to the functional $\mathcal{R}_\kappa$, seen as a function of the i.i.d. random variables $(\varepsilon_l, X_{(l-1)d_1+1}^{(1)}, \ldots, X_{ld_1}^{(1)}, \ldots, X_{(l-1)d_K+1}^{(K)}, \ldots, X_{ld_K}^{(K)})$, $1 \leq l \leq \kappa$: changing any of these random variables change the value of $\mathcal{R}_\kappa$ by at most $\mathcal{M}_{\mathcal{H}}/\kappa$. One thus obtains from (3.8) with $\psi(x) = \exp(\lambda x)$, where $\lambda > 0$ is a parameter which shall be chosen later, that:

$$\mathbb{E} \left[\exp \left(\lambda \sup_{H \in \mathcal{H}} \left| U_{\mathbf{n}}(\bar{H}) \right| \right) \right] \leq \exp \left(2\lambda \mathbb{E}[\mathcal{R}_\kappa] + \frac{\mathcal{M}_{\mathcal{H}}^2 \lambda^2}{4\kappa} \right). \tag{3.9}$$

Applying Chernoff's method, one then gets:

$$\mathbb{P}\left\{\sup_{H\in\mathcal{H}}\left|U_{\mathbf{n}}(\bar{H})\right|>\eta\right\}\leq\exp\left(-\lambda\eta+2\lambda\mathbb{E}[\mathcal{R}_{\kappa}]+\frac{\mathcal{M}_{\mathcal{H}}^{2}\lambda^{2}}{4\kappa}\right). \tag{3.10}$$

Using the bound (see Eq. (6) in [Boucheron *et al.*, 2005] for instance)

$$\mathbb{E}[\mathcal{R}_{\kappa}]\leq\mathcal{M}_{\mathcal{H}}\sqrt{\frac{2V\log(1+\kappa)}{\kappa}}$$

and taking $\lambda=2\kappa(\eta-2\mathbb{E}[\mathcal{R}_{\kappa}])/\mathcal{M}_{\mathcal{H}}^{2}$ in (3.10) finally establishes the desired result.

Part II

Algorithms for K-partite ranking

Chapter 4

Aggregation of scoring functions

The purpose of this chapter is to build consistent scoring functions for multipartite ranking problem using bipartite scoring functions i.e. scoring functions learned using only two labels. Indeed, a natural approach is to transfer virtuous bipartite ranking methods to derive optimal and consistent rules for K-partite ranking. This idea is quite successful in the multiclass classification setup (see [Hastie & Tibshirani, 1998] or [Fürnkranz, 2002] for instance). We propose to build on the original proposition in [Fürnkranz *et al.*, 2009] to combine bipartite ranking tasks in order to solve the K-partite case.

The first step is to decompose the multipartite ranking problem into a set of bipartite ranking problems. In ordinal regression, it is shown that decompositions that take into account the order on the set of the labels are preferable, see [Frank & Hall, 2001], [Fürnkranz *et al.*, 2009]. So we use the same decompositions for the multipartite ranking. Once the decomposition is chosen we learn a scoring function for each sub-problem.

The second step consists in aggregating these scoring functions. This step is more complicated than calculating a mean since what we want to aggregate are the orders induced by the scoring functions. However, it is possible to define metrics on the order induced by scoring functions such as the Kendall τ distance. The aggregation step consists in choosing the scoring function that minimizes the sum of the Kendall τ distances with the learned scoring functions. This function establishes a ranking consensus, called a *median scoring function*. It is also shown that such a median scoring function always exists in the important situation where the scoring functions one seeks to summarize/aggregate are piecewise constant, and computation of this median function is feasible. We call the final procedure the Kendall aggregation for multipartite ranking.

Then we study the consistency of the aggregation procedure in two scenario. The easy case is when all the supports of the conditional densities are the same. In this situation, it is possible to link the deficit of AUC and the Kendall τ distance. Using this relationship, we state that under the monotone likelihood ratio condition together with a margin condition over the posterior distributions, the median scoring function built out of pairwise AUC-consistent function is VUS-consistent. However, in practice supports are rarely the same and an issue called the supports issue can happen. Basically, the problem comes down that the structure of one bipartite subproblem is not the same as the multipartite task. Using the decomposition step introduced in [Frank & Hall, 2001], we show that the procedure is VUS-consistent

if the scoring functions are consistent for a certain quantity that looks like the AUC.

The rest of the chapter is organized as follows. In the section 4.1, we present the Kendall τ metric and the median procedure. In the section 4.2, we introduce the VUS-consistency and the low noise assumption that are required to state the theorem. We also present the links between the deficit of AUC and the Kendall τ distance. Finally, we state the consistency of the Kendall aggregation procedure using the decomposition [Frank & Hall, 2001] when the supports are not the same.

4.1 Pairwise aggregation: from bipartite to K-partite ranking

In the present section, we propose a practical strategy for building scoring functions which approximate optimal scoring functions for multipartite ranking based on a set of labeled observations. The principle of this strategy is the aggregation of scoring functions obtained for the pairwise subproblems. We emphasize the fact that the situation is very different from multiclass classification where aggregation boils down to linear combination, or majority voting, over binary classifiers (for "*one against one*" and "*one versus all*", we refer to [Allwein *et al.*, 2001], [Hastie & Tibshirani, 1998], [Venkatesan & Amit, 1999], [Debnath *et al.*, 2004], [Dietterich & Bakiri, 1995], [Beygelzimer *et al.*, 2005b], [Beygelzimer *et al.*, 2005a] and the references therein for instance). We propose here, in the K-partite ranking setup, a metric-based barycentric approach to build the aggregate scoring function from the collection of scoring functions estimated for the bipartite subproblems.

4.1.1 Decomposition step

The first stage consists in decomposing the K-partite ranking problem in a collection of bipartite ranking tasks. We review here the existing decomposition methods and present some other that can be used in the case of ordinal label. The first method was initially proposed to solve the problem of ordinal regression by Frank and Hall [Frank & Hall, 2001] so we call it (FH). It consists in learning $(K-1)$ bipartite ranking functions, specifically the observations with labels less than k against the labels strictly greater than k, for $k \in \{1, \ldots, K-1\}$ (i.e $\cup_{i=1}^{k} \mathcal{D}_i$ vs $\cup_{i=k+1}^{K} \mathcal{D}_i$). In [Frank & Hall, 2001], each of the problems provides an estimate of the function $\mathbb{P}\{Y > k | X = x\}$ which allows us to find an estimator of $\eta(x) = \mathbb{E}[Y|X = x]$, summing up the estimated functions. The second method proposed in [Fürnkranz *et al.*, 2009] is called round-robin or learning by pairwise comparison (RR) and consists in learning a ranking function for each pair of labels (i.e $\mathcal{D}_k$ vs $\mathcal{D}_l$ for all $k < l$). This gives $K(K-1)/\,2$ ranking functions, which they aggregated by summing the values of the scoring functions. Using several benchmark dataset, they show that this method outperforms SVM with linear kernels.

Many strategies can be considered and we introduce two decompositions here that we use in the sequel. Using the upper bound of the deficit of VUS in function

of the deficit of AUC (see Theorem 1.2.12), it appears that a scoring function that is good for the problem 1 vs 2 and 2 vs 3 is good for the multipartite ranking problem. So it is natural to consider a scoring function that respect the most the orders induced by these two scoring functions. In this case summing the values is not the best strategy to aggregate and the next subsection explains how to create a scoring function that induces an order close to the ones induced by the two learned function. Finally, in experiments, we use a decomposition that is intermediate between the two first approaches. We solve $K(K-1)/2$ ranking problems, each one corresponding to the observations with a label smaller than k against the observations that have a label greater than l (i.e. $\cup_{i=1}^{k}\mathcal{D}_i$ vs. $\cup_{i=l}^{K}\mathcal{D}_i$), for $k < l$. Of course all the problems of decomposition FH are included in this decomposition and for this reason we call this decomposition pairwise FH (PFH).

Remark 4.1.1. *In classification, there exists a popular method called "one versus all" (OVA) that consists in estimating the probabilities $\eta_k(x) = \mathbb{P}\{Y = k|X = x\}$ by solving the problem "k" against all other classes (i.e. $\mathcal{D}_k$ vs. $\cup_{j\neq k}^{K}\mathcal{D}_j$). From these estimates, we can create a scoring function which estimates the regression function $\eta(x) = \sum_{k=1}^{K} k\eta_k(x)$. However, we can not use this decomposition to learn scoring functions that we aggregate in a consensus ranking rule because the induced orders are not consistent. This decomposition does not take into account the orderliness of labels and several experimental results shows that this method is less effective in such a context [Huhn & Hüllermeier, 2009].*

4.1.2 Median scoring functions and optimal aggregation

Every scoring function induces an order relation over the input space $\mathbb{R}^d$ and, for the ranking problem considered here, a measure of similarity between two scoring functions should only take into consideration the similarity in the ranking induced by each one of them. We propose here a measure of agreement between scoring functions which is based on the probabilistic Kendall τ for a pair of random variables.

Definition 4.1.1. (Probabilistic Kendall τ) *Consider X, X' i.i.d. random vectors with density function ϕ over $\mathbb{R}^d$. The measure of agreement between two real-valued scoring functions s_1 and s_2 is defined as the quantity:*

$$\begin{aligned}\tau(s_1, s_2) = \mathbb{P}\left\{\big(s_1(X) - s_1(X')\big)\cdot\big(s_2(X) - s_2(X')\big) > 0\right\} \\ + \frac{1}{2}\mathbb{P}\left\{s_1(X) \neq s_1(X'),\ s_2(X) = s_2(X')\right\} \\ + \frac{1}{2}\mathbb{P}\left\{s_1(X) = s_1(X'),\ s_2(X) \neq s_2(X')\right\}\ .\end{aligned}$$

This definition of agreement between scoring functions s_1 and s_2 coincides indeed with the Kendall τ between real-valued random variables $s_1(X)$ and $s_2(X)$. Note that the contribution of the two last terms in the definition of $\tau(s_1, s_2)$ vanishes

when the distributions of the $s_i(X)$'s are continuous. Moreover, this quantity is very similar to the AUC and one can easily see that

$$\tau(s(X), Y) = 2p(1-p)\mathrm{AUC(s)} + 1/2 \cdot \mathbb{P}\{\mathrm{s(X) \neq s(X')}, \ \ \mathrm{Y = Y'}\},$$

where $Y \in \{1, 2\}$. This similarity plays a crucial role when proving the consistency theorems (see 4.2).

Then one can define the notion of median scoring function which accounts for the consensus of many real-valued scoring functions over a given class of candidates.

Definition 4.1.2. (Median scoring function) *Consider a given class $\mathcal{S}_1$ of real-valued scoring functions and $\Sigma_K = \{s_1, \ldots, s_{K-1}\}$ a finite set of real-valued scoring functions. A median scoring function $\overline{s}$ for $(\mathcal{S}_1, \Sigma_K)$ satisfies:*

$$\sum_{k=1}^{K-1} \tau\left(\overline{s}, s_k\right) = \sup_{s \in \mathcal{S}_1} \sum_{k=1}^{K-1} \tau\left(s, s_k\right). \tag{4.1}$$

In general, the supremum appearing on the right hand side of Eq. (4.1) is not attained. However, when the supremum over $\mathcal{S}_1$ can be replaced by a maximum over a finite set $\mathcal{S}_1' \subset \mathcal{S}_1$, a median scoring function always exists (but it is not necessarily unique). In particular, this is the case when considering *piecewise constant scoring functions* such as those produced by the bipartite ranking algorithms proposed in [Clémençon *et al.*, 2011a], [Clémençon & Vayatis, 2010], [Clémençon & Vayatis, 2009a] (we also refer to [Clémençon & N.Vayatis, 2009] for a discussion of consensus computation/approximation in this case). The idea underlying the measure of consensus through Kendall metric in order to aggregate scoring functions that are nearly optimal for bipartite ranking subproblems is clarified by the following result.

Definition 4.1.3. (Pairwise optimal scoring function) *A pairwise optimal scoring function $s_{l,k}^*$ is an optimal scoring function for the bipartite ranking problem with classes $Y = k$ and $Y = l$, where $k > l$ in the sense that:*

$$\forall x, x' \in \mathcal{X} \ , \quad \Phi_{k,l}(x) < \Phi_{k,l}(x') \Rightarrow s_{l,k}^*(x) < s_{l,k}^*(x') \ .$$

We denote by $\mathcal{S}_{l,k}^*$ the set of such optimal functions and, in particular, $\mathcal{S}_k^* = \mathcal{S}_{k,k+1}^*$.

Proposition 4.1.1. *Denote by $\mathcal{S}$ the set of all possible real-valued scoring functions and consider pairwise optimal scoring functions $s_k^* \in \mathcal{S}_k^*$ for $k = 1, \ldots, K-1$, which form the set $\Sigma_K^* = \{s_1^*, \ldots, s_{K-1}^*\}$. Under Assumption 1, we have:*

1. *A median scoring function $\overline{s}^*$ for $(\mathcal{S}, \Sigma_K^*)$ is an optimal scoring function for the K-partite ranking problem.*

2. *Any optimal scoring function s^* for the K-partite ranking problem satisfies:*

$$\sum_{k=1}^{K-1} \tau(s^*, s_k^*) = K - 1 \ .$$

The proposition above reveals that "consensus scoring functions", in the sense of Definition 4.1.2, based on $K-1$ optimal scoring functions are still optimal solutions for the global K-partite ranking problem and that, conversely, optimal elements necessarily achieve the equality in Statement (2) of the previous proposition. This naturally suggests to implement the following two-stage procedure, that consists in 1) solving the bipartite ranking subproblem related to the pairwise case $(k, k+1)$ of consecutive class labels, yielding a scoring function s_k, for $1 \leq k < K$, and 2) computing a median according to Definition 4.1.2, when feasible, based on the latter over a set $\mathcal{S}_1$ of scoring functions. Beyond the difficulty to solve each ranking subproblem separately (for instance refer to [Clémençon & Vayatis, 2009b] for a discussion of the nature of the bipartite ranking issue), the performance/complexity of the method sketched above is ruled by the richness of the class $\mathcal{S}_1$ of scoring function candidates: too complex classes clearly make median computation unfeasible, while poor classes may not contain sufficiently accurate scoring functions.

4.1.3 A practical aggregation procedure

We now propose to convert the previous theoretical results which relate pairwise optimality to K-partite optimality in ranking into a practical aggregation procedure. Consider two independent samples:

- a sample $\mathcal{D} = \{(X_i, Y_i) : \ 1 \leq i \leq n\}$ with i.i.d. labeled observations,
- a sample $\mathcal{D}' = \{X'_i, : \ 1 \leq i \leq n'\}$ a sample with unlabeled observations.

The first sample $\mathcal{D}$ is used for training bipartite ranking functions $\hat{s}_k$, while the second sample $\mathcal{D}'$ will be used for the computation of the median. In practice a proxy for the median is computed based on the empirical version of the Kendall τ, the following U-statistic of degree two, see [Clémençon *et al.*, 2008].

Definition 4.1.4. (Empirical Kendall τ) *Given a sample $X_1, \ldots, X_n$, the empirical Kendall τ is given by:*

$$\widehat{\tau}_n(s_1, s_2) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} h\left((s_1(X_i), s_1(X_j)), (s_2(X_i), s_2(X_j))\right) \ ,$$

where

$$h\left((v, w), (v', w')\right) = \mathbb{I}\left\{\left(v - v'\right) \cdot \left(w - w'\right) > 0\right\} + \frac{1}{2}\mathbb{I}\left\{v = v', \ w \neq w'\right\} + \frac{1}{2}\mathbb{I}\left\{v \neq v', \ w = w'\right\},$$

for (v, w) and (v', w') in $\mathbb{R}^2$.

The following aggregation method describes a two-steps procedure which takes as input the two data sets, a class $\mathcal{S}_1$ of candidate scoring functions, and a generic bipartite ranking algorithm $\mathcal{A}$.

KENDALL AGGREGATION FOR K-PARTITE RANKING

Input. Data samples $\mathcal{D}$ and $\mathcal{D}'$, a bipartite ranking algorithm $\mathcal{A}$, a class $\mathcal{S}_1$ of scoring functions.

1. **Build pairwise scoring functions for bipartite ranking.** For $k = 1, \ldots, K-1$, run algorithm $\mathcal{A}$ in order to train a scoring function $\widehat{s}_k$ based on the restricted samples $\mathcal{D}_k \cup \mathcal{D}_{k+1}$.

2. **Aggregate pairwise scoring functions for K-partite ranking.** Compute

$$\widehat{s} = \operatorname*{arg\,max}_{s \in \mathcal{S}_1} \sum_{k=1}^{K-1} \widehat{\tau}' \left(s, \widehat{s}_k \right),$$

where $\widehat{\tau}'$ is the empirical Kendall τ computed over the sample $\mathcal{D}'$.

Output. Empirical median scoring function $\widehat{s}$ in $\mathcal{S}_1$ for K-partite ranking

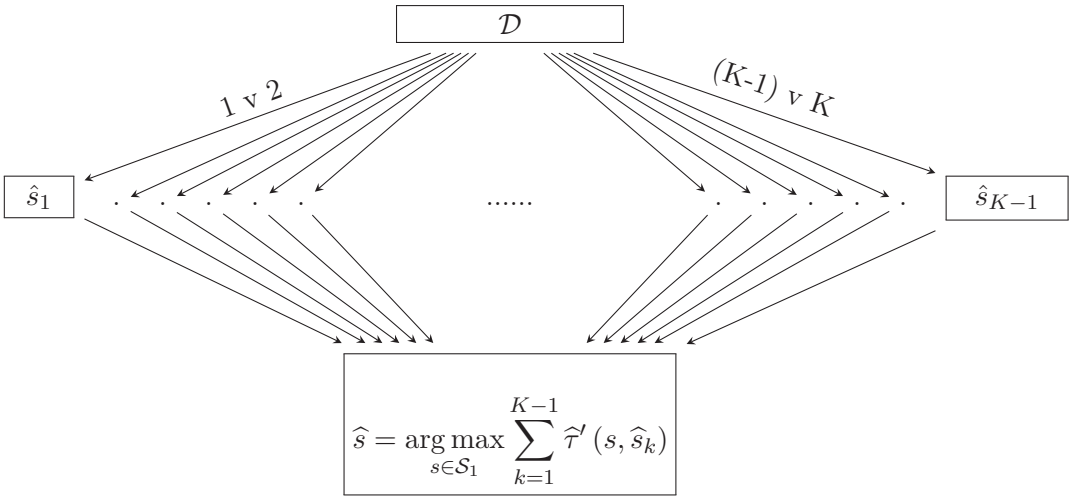


Figure 4.1: Kendall-τ aggregation procedure.

4.1.3.1 Practical implementation issues.

Motivated by practical problems such as the design of meta-search engines, collaborative filtering or combining results from multiple databases, *consensus ranking*, which the second stage of the procedure described above is a special case of, has recently enjoyed renewed popularity and received much attention in the machine-learning literature, see [Meila *et al.*, 2007], [Fagin *et al.*, 2003] or [Lebanon & Lafferty, 2003] for instance. As shown in [Hudry, 2008] or [Wakabayashi, 1998] in particular, median computations are NP-hard problems in general. Except in the case where $\mathcal{S}_1$ is of very low cardinality, the (approximate) computation of a supremum involves in practice the use of meta-heuristics such as simulated annealing, tabu search or genetic algorithms. The description of these computational approaches to consensus ranking is beyond the scope of this thesis and we refer to [Barthélemy *et al.*, 1989], [Charon & Hudry, 1998], [Laguna *et al.*, 1999] or [Mandhani & Meila, 2009] and the references therein for further details on their implementation. We also underline that the procedure of the Kendall aggregation approach is described using the decomposition k vs $k+1$ but could be replaced by any of the decomposition described in 4.1.1.

Rank prediction vs. scoring function learning. When the goal is to rank accurately new unlabeled datasets, rather than to learn a nearly optimal scoring function explicitly, the following variant of the procedure described above can be considered. Given an unlabeled sample of i.i.d. copies of the input r.v. X $\mathcal{D}_X = \{X_1, \ldots, X_m\}$, instead of aggregating scoring functions s_k defined on the feature space $\mathcal{X}$ and use a consensus function for ranking the elements of $\mathcal{D}_X$, one may aggregate their restrictions to the finite set $\mathcal{D}_X \subset \mathcal{X}$, or simply the ranks of the unlabeled data as defined by the s_k's.

4.2 Consistency of pairwise aggregation

In this section, we assume a data sample $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ is available and composed by n i.i.d. copies of the random pair (X, Y). Our goal here is to learn from the sample $\mathcal{D}_n$ how to build a real-valued scoring function $\widehat{s}_n$ such that its ROC surface is as close as possible to the optimal ROC surface.

4.2.1 Definition of VUS-consistency and main result

We need to use the notion of AUC consistency for the bipartite ranking subproblems.

Definition 4.2.1. (AUC CONSISTENCY) *For k fixed in $\{1, \ldots, K-1\}$, a sequence $(s_n)_{n\geq 1}$ of scoring functions is said to be AUC-consistent (respectively, strongly AUC-consistent) for the bipartite problem (ϕ_k, ϕ_{k+1}) if it satisfies:*

$$\mathrm{AUC}_{\phi_k,\phi_{k+1}}(s_n) \to \mathrm{AUC}^*_{\phi_k,\phi_{k+1}} \text{ in probability (resp., with probability one).}$$

We propose to consider a weak concept of consistency which relies on the VUS.

Definition 4.2.2. (VUS-CONSISTENCY) *Suppose that Assumption 1 is fulfilled. Let $(s_n)_{n\geq 1}$ be a sequence of random scoring functions on $\mathbb{R}^d$, then:*

- *the sequence $\{s_n\}$ is called* VUS*-consistent if*

$$\mathrm{VUS}^* - \mathrm{VUS}(\mathrm{s_n}) \to 0 \quad \textit{in probability,}$$

- *the sequence $\{s_n\}$ is called strongly* VUS*-consistent if*

$$\mathrm{VUS}^* - \mathrm{VUS}(\mathrm{s_n}) \to 0 \quad \textit{with probability one.}$$

Remark 4.2.1. *We note that the deficit of VUS can be interpreted as an L_1 distance between ROC surfaces of s_n and s^*:*

$$\mathrm{VUS}^* - \mathrm{VUS}(\mathrm{s_n}) = \int\int_{(\alpha,\gamma)\in[0,1]^2} |\mathrm{ROC}(\mathrm{s}^*,\alpha,\gamma) - \mathrm{ROC}(\mathrm{s_n},\alpha,\gamma)|\ \mathrm{d}\alpha\ \mathrm{d}\gamma\ ,$$

and in this sense the notion of consistency is weak. Indeed, a stronger sense of consistency could be given by considering the supremum norm between surfaces:

$$d_\infty(s^*, s_n) = \sup_{(\alpha,\gamma)\in[0,1]^2} |\mathrm{ROC}(\mathrm{s}^*,\alpha,\gamma) - \mathrm{ROC}(\mathrm{s_n},\alpha,\gamma)|,$$

The study of accuracy of K-partite ranking methods in this sense is beyond the scope of the present chapter (in contrast to the L_1 norm, the quantity $d_\infty(s^*, s)$ cannot be decomposed in an additive manner). Extensions of bipartite ranking procedures such as the TreeRank *and the* RankOver *algorithms (see [Clémençon & Vayatis, 2009b] and [Clémençon & Vayatis, 2010]), for which consistency in supremum norm is guaranteed under some specific assumptions, will be considered in Chapter 5.*

In order to state the main result, we need an additional assumption on the distribution of the random pair (X, Y). The reason why this assumption is needed will be explained in the next section.

Assumption 2. *For all $k \in \{1, \ldots, K-1\}$, the (pairwise) posterior probability given by $\eta_{k+1}(X)/(\eta_k(X) + \eta_{k+1}(X))$ is a continuous random variable and there exist $c < \infty$ and $a \in (0, 1)$ such that*

$$\forall x \in \mathcal{X},\ \ \mathbb{E}\left[\left|\frac{\eta_{k+1}(X)}{\eta_{k+1}(X)+\eta_k(X)} - \frac{\eta_{k+1}(x)}{\eta_{k+1}(x)+\eta_k(x)}\right|^{-a}\right] \leq c\ . \tag{4.2}$$

This hypothesis measures how much data are spread and quantifies the difficulty of the ranking task. If a is near 0 then the data are very concentrated and it is difficult to order, if a is close to 1 observations are spread between 0 and 1. In

the statistical learning literature, Assumption 2 is referred to as the noise condition and goes back to the work of Tsybakov [Tsybakov, 2004]. It was adapted to the framework of bipartite ranking in [Clémençon *et al.*, 2008]. This type of low noise conditions is deeply studied in the last chapter of this manuscript (chapter 7).

We can now state the main consistency result of the paper which concerns the Kendall aggregation procedure described in Section 4.1.3. Indeed, the following theorem reveals that the notion of median scoring function introduced in Definition 4.1.2 preserves AUC consistency for bipartite subproblems and thus yields a VUS consistent scoring function for the K-partite problem. It is assumed that the solutions to the bipartite subproblems are AUC-consistent for each specific pair of class distributions (ϕ_k, ϕ_{k+1}), $1 \leq k < K$. For simplicity, we formulate the result in the case $K = 3$.

Theorem 4.2.1. *We consider a class of candidate scoring functions $\mathcal{S}_1$, $(s_n^{(1)})_{n\geq 1}$, $(s_n^{(2)})_{n\geq 1}$ two sequences of scoring functions in $\mathcal{S}_1$. We use the notation $\Sigma_{2,n} = \{s_n^{(1)}, s_n^{(2)}\}$. Assume the following:*

1. *Assumptions 1 and 2 hold true.*
2. *The class $\mathcal{S}_1$ contains an optimal scoring function.*
3. *The sequences $(s_n^{(1)})_{n\geq 1}$ and $(s_n^{(2)})_{n\geq 1}$ are (strongly)* AUC*-consistent for the bipartite ranking subproblems related to the pairs of distributions (ϕ_1, ϕ_2) and (ϕ_2, ϕ_3) respectively.*
4. *Assume that, for all n, there exists a median scoring function $\overline{s}_n$ in the sense of Definition 4.1.2 with respect to $(\mathcal{S}_1, \Sigma_{2,n})$.*

Then the median scoring function $\overline{s}_n$ is (strongly) VUS*-consistent.*

Discussion. The first assumption of theorem 4.2.1 puts a restriction on the class of distributions for which such a consistency result holds. Assumption 1 actually guarantees that the very problem of K-partite makes sense and the existence of an optimal scoring function. Assumption 2 can be seen as a "light" restriction since it still covers a large class of distributions commonly used in probabilistic modeling. The third and fourth assumptions are natural as we expect first to have efficient solutions to the bipartite subproblems before considering reasonable solutions to the K-partite problem. The most restrictive assumption is definitely the second one about the fact that the class of candidates contains an optimal element. Indeed, it is easy to weaken this assumption at the price of an additional bias term by assuming that the scoring functions $s_n^{(1)}$, $s_n^{(2)}$ and $\overline{s}_n$ belong to a set $\mathcal{S}_1^{(n)}$, such that there exists a sequence $(s_n^*)_{n\geq 1}$ with $s_n^* \in \mathcal{S}_1^{(n)}$ and $\mathrm{VUS}(s_n^*) \to \mathrm{VUS}^*$ as $n \to \infty$. We decided not to include this refinement as this is merely a technical argument which does not offer additional insights on the nature of the problem.

4.2.2 From AUC consistency to VUS consistency

In this section, we introduce auxiliary results which contribute to the proof of the main theorem (details are provided in the last section of this chapter). Key arguments rely on the relationship between the solutions of the bipartite ranking subproblems and those of the K-partite problem. In particular, a sequence of scoring functions that is simultaneously AUC-consistent for the bipartite ranking problems related to the two pairs of distributions (ϕ_1, ϕ_2) and (ϕ_2, ϕ_3) is VUS-consistent. Indeed, we have the following corollary.

Corollary 4.2.2. *Suppose that Assumption 1 is satisfied. Let $(s_n)_{n\geq 1}$ be a sequence of scoring functions. The following assertions are equivalent.*

(i) The sequence $(s_n)_n$ of scoring functions is (strongly) VUS-optimal.

(ii) We have simultaneously when $n \to \infty$:

$$\begin{aligned} \mathrm{AUC}_{\phi_1,\phi_2}(\mathrm{s_n}) &\to \mathrm{AUC}^*_{\phi_1,\phi_2} \\ \mathrm{AUC}_{\phi_2,\phi_3}(\mathrm{s_n}) &\to \mathrm{AUC}^*_{\phi_2,\phi_3} \ . \end{aligned}$$

(with probability one) in probability.

It follows from this result that the 3-partite ranking problem can be cast in terms of a double-criterion optimization task, consisting in finding a scoring function s that simultaneously maximizes $\mathrm{AUC}_{\phi_1,\phi_2}(\mathrm{s})$ and $\mathrm{AUC}_{\phi_2,\phi_3}(\mathrm{s})$. This result provides a theoretical basis for the justification of our pairwise aggregation procedure.

The other type of result which is needed concerns the connection between the aggregation principle based on a consensus approach (Kendall τ) and the performance metrics involved in the K-partite ranking problem. The next results establish inequalities which relate the AUC and the Kendall τ in a quantitative manner.

Proposition 4.2.3. *Let p be a real number in $(0, 1)$. Consider two probability distributions ϕ_k and $\phi_k + 1$ on the set $\mathcal{X}$. We assume that the distribution of X comes from the mixture with density function given by $(1-p)\phi_k + p\phi_{k+1}$. For any real-valued scoring functions s_1 and s_2 on $\mathbb{R}^d$, we have:*

$$|\mathrm{AUC}_{\phi_\mathrm{k},\phi_{\mathrm{k}+1}}(\mathrm{s}_1) - \mathrm{AUC}_{\phi_\mathrm{k},\phi_{\mathrm{k}+1}}(\mathrm{s}_2)| \leq \frac{1 - \tau\,(\mathrm{s}_1, \mathrm{s}_2)}{4\mathrm{p}(1-\mathrm{p})} \ .$$

We point out that it is generally vain to look for a reverse control: indeed, scoring functions yielding different rankings may have exactly the same AUC. However, the following result guarantees that a scoring function with a nearly optimal AUC is close to optimal scoring functions in a certain sense, under the additional assumption that the noise condition introduced in [Clémençon *et al.*, 2008] is fulfilled.

Proposition 4.2.4. *Under Assumption 2, we have, for any $k \in \{1, \dots, K-1\}$, for any scoring function s and any pairwise optimal scoring function s_k^*:*

$$1 - \tau\left(s_k^*, s\right) \leq C \cdot \left(\mathrm{AUC}^*_{\phi_k, \phi_{k+1}} - \mathrm{AUC}_{\phi_k, \phi_{k+1}}(s)\right)^{a/(1+a)},$$

with $C = 3c^{1/(1+a)} \cdot (2p_k p_{k+1})^{a/(1+a)}$.

4.3 How to solve the supports issue

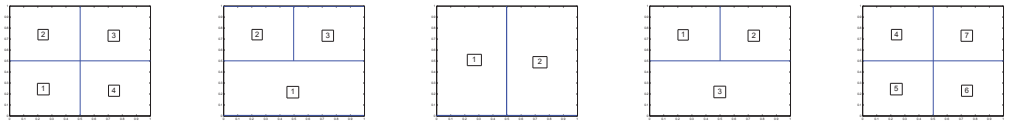
In the previous section, the supports of the conditional density are the same. However, when the supports are not all equal, some issues come down. First, we explain with a toy example what we call the supports issue and we state a consistency theorem using the decomposition (FH) combined with the Kendall aggregation.

4.3.1 The supports issue

The supports issue appears when we decompose the multi-class problem into a series of bipartite problems. Indeed, it may happen that, when considering one of the bipartite problems, a part of the space is not taken into account because there is no data in the training set. We illustrate this by a toy example.
We consider the case where $K = 3$ and the data follow the following distribution: cut the unit square $[0,1]^2$ in four areas $A_1 = [0, 1/2] \times [0, 1/2]$, $A_2 = [0, 1/2] \times [1/2, 1]$, $A_3 = [1/2, 1] \times [1/2, 1]$, $A_4 = [1/2, 1] \times [0, 1/2]$ take μ the uniform distribution on the square and $\eta_1(x) = 1\mathbb{I}\{X \in A_1\} + (1/12)\mathbb{I}\{X \in A_2\}$, $\eta_2(x) = (11/12)\mathbb{I}\{X \in A_2\} + (11/12)\mathbb{I}\{X \in A_3\}$ and $\eta_3(x) = (1/12)\mathbb{I}\{X \in A_3\} + 1\mathbb{I}\{X \in A_4\}$. The optimal function is given in Figure 4.2) a. Suppose we want to find a consensus scoring function from the decomposition 1vs2, 2vs3 and 1vs3 where each problem is treated by a binary tree. The optimal function for each problem and the consensus scoring function based on the average ranks are shown in Figure 4.2b.c.d.e. Here, it is possible to compute the scoring function that maximizes the sum of the Kendall-τ and we find that it is the scoring function "1vs3" and the sum equals 2.375. In each bipartite problem of the FH decomposition, all the data are used whereas for the RR decomposition only two labels are considered in each problem. So we expect that the FH decomposition is less sensitive to supports issue than the RR decomposition. We highlight this from a theoretical angle in the next section and with experiments in chapter 6.

4.3.2 Consistency even with supports issue

The purpose of this section is to prove a theorem of consistency in terms of VUS in the case where $K = 3$. We study the Kendall aggregation approach for multipartite ranking described in section 4.1. What is paramount in this study is the relationship between Kendall-τ and loss related to the problem $\mathcal{D}_1$ vs $\mathcal{D}_2 \cup \mathcal{D}_3$ and $\mathcal{D}_1 \cup \mathcal{D}_2$ vs



a. Optimal scoring function | b. Scoring function 1vs2 | c. Scoring function 1vs3 | d. Scoring function 2vs3 | e. Consensus scoring function

Figure 4.2: Scoring functions for the toy example

$\mathcal{D}_3$. The following two propositions will clarify these links. First, let $M_{1v23}(s) = \mathbb{P}\{s(X) < s(X')|Y = 1, Y' \in \{2,3\}\} + \frac{1}{2}\mathbb{P}\{s(X) = s(X')|Y = 1, Y' \in \{2,3\}\}$ be the probability to well-order a pair of observations given that one observation has the label 1 and the other has the label 2 or 3. Similarly, we note $M_{12v3}(s) = \mathbb{P}\{s(X) < s(X')|Y \in \{1,2\}, Y' = 3\} + \frac{1}{2}\mathbb{P}\{s(X) = s(X')|Y \in \{1,2\}, Y' = 3\}$. We say that the sequence of scoring functions s_n is M-consistent for the task $\mathcal{D}_1$ vs $\mathcal{D}_2 \cup \mathcal{D}_3$ if $M_{1v23}(s_n)$ tends to M^*_{1v23} when the number of observations tends to infinity, where M^*_{1v23} refers to $\max_{s \in \mathcal{S}} M_{1v23}(s)$. It is easy to check that $M_{1v23}(s^*) = M^*_{1v23}$, for all $s^* \in \mathcal{S}^*$. In particular, we have the following lemma.

Lemma 4.3.1. *For all $s \in \mathcal{S}$*

$$M^*_{1v23} - M_{1v23}(s) = \frac{1}{p_1(1-p_1)}\mathbb{E}[|\eta_1(X) - \eta_1(X')|\mathbb{I}\{\Gamma_{s,1}(X, X')\}] + \frac{1}{p_1(1-p_1)}\mathbb{P}\{s(X) = s(X')|Y = 1, Y' \in \{2,3\}\}$$

$$M^*_{12v3} - M_{12v3}(s) = \frac{1}{p_3(1-p_3)}\mathbb{E}[|\eta_3(X) - \eta_3(X')|\mathbb{I}\{\Gamma_{s,3}(X, X')\}] + \frac{1}{p_3(1-p_3)}\mathbb{P}\{s(X) = s(X')|Y \in \{1,2\}, Y' = 3\}$$

where $\Gamma_{s,i} = \{(x, x') \in \mathcal{X} \times \mathcal{X} | (s(x) - s(x'))(\eta_i(x) - \eta_i(x')) < 0\}$

With this lemma, we obtain the following proposition

Proposition 4.3.2. *For all $s_1, s_2 \in \mathcal{S}$,*

$$|M_{1v23}(s_1) - M_{1v23}(s_2)| \leq \frac{1 - \tau(s_1, s_2)}{p_1(1-p_1)},$$

$$|M_{12v3}(s_1) - M_{12v3}(s_2)| \leq \frac{1 - \tau(s_1, s_2)}{p_3(1-p_3)}$$

where $\mu(dx) = p_1F_1(dx) + p_2F_2(dx) + p_3F_3(dx)$.

This proposition allows us to bound the difference in gain by the Kendall-τ distance between two scoring functions. Note that this result is valid for any pair of scoring functions and requires no assumption about the distribution P. It is used to bound the gap between $M_{1v23}(s)$ and M^*_{1v23}. To show the consistency of the procedure, we need an inequality in the other direction, i.e., we want to bound the Kendall-τ distance by the deficit in gain between s and s^*. Assumptions on (η_1, η_2, η_3) the posterior distributions are needed

Proposition 4.3.3. *Assume that the assumption 1 and 2 hold. Then, we have for all pair of real valued scoring functions* $(s, s^*) \in \mathcal{S} \times \mathcal{S}^*$,

$$1 - \tau(s^*, s) \leq C_1 \cdot (M^*_{1v23} - M_{1v23}(s))^{a/(1+a)},$$

$$1 - \tau(s^*, s) \leq C_2 \cdot (M^*_{12v3} - M_{12v3}(s))^{a/(1+a)},$$

with $C_1 = c^{1/(1+a)}(2p_1(1-p_1))^{a/(1+a)}$ *and* $C_2 = c^{1/(1+a)}(2p_3(1-p_3))^{a/(1+a)}$.

Finally, we can state the theorem establishing the consistency of the procedure of breaking down with FH method and aggregating using the Kendall-τ distances.

Theorem 4.3.4. *Suppose that assumptions of Proposition 4.3.3 are satisfied. Let* $\mathcal{S}_1 \subset \mathcal{S}_0$ *be some set of real-valued scoring functions such that* $\mathcal{S}^* \cap \mathcal{S}_1 \neq \emptyset$. *Let* $s_n(x)$ *and* $s'_n(x)$ *be* M*-consistent sequences of scoring functions in* $\mathcal{S}_1$ *for the bipartite ranking problems related to the pairs of distributions* $1v23$ *and* $12v3$ *respectively. If there exists a median scoring function* $\overline{s}_n(x)$ *in the sense of Definition 4.1.2, then it is* VUS*-consistent.*

It is essential to note that to show the consistency, we make no assumption on the supports of the observations. From a theoretical point of view, we see that the supports issue can be solved using the decomposition FH.

4.4 Conclusion

In this chapter, we present methods to solve the multipartite ranking problem. The first part is dedicated to describe the methods for decomposing the multipartite ranking problem into a series of bipartite ranking tasks, as proposed in [Fürnkranz *et al.*, 2009]. We have introduced a specific notion of *median scoring function* based on the (probabilistic) Kendall τ distance. When the supports of the conditionnal densities are equal, it is shown that the aggregation procedure leads to a consistent ranking rule, when applied to scoring functions that are, each, consistent for the bipartite ranking subproblem related to a specific pair of consecutive class distributions. This approach allows for extending the use of ranking algorithms originally designed for the bipartite situation to the ordinal multi-class context. We highlight, thanks to an example, that a decomposition can create inconsistencies when the supports of the observations is not the same given the label. However, it is shown that the decomposition proposed by Frank and Hall [Frank & Hall, 2001]

coupled with the Kendall aggregation produces a consistent ranking rule when each of the scoring functions is consistent for their problem. This result is true without any assumption on the supports of the observations.

Finally, we underline that, so far, very few practical algorithms tailored for ROC graph optimization have been proposed in the literature. Whereas, as shown at length in [Clémençon & Vayatis, 2009b] and [Clémençon *et al.*, 2011a], partitioning techniques for AUC maximization, in the spirit of the CART method for classification, can be implemented in a simple manner, by solving recursively cost-sensitive classification problems (with a local cost, depending on the data lying in the cell to be split), recursive VUS maximization remains a challenging issue, for which no simple interpretation is currently available. It is the purpose of the next chapter to study this problem.

4.5 Proofs

Proof of Proposition 4.1.1

Under Assumption 1, the regression function η is an optimal scoring function (see Theorem 1.1.1 (3)). Using the fact that $\phi_{k+1,k} \in \mathcal{S}^*_{k+1,k}$ combined with Theorem 1.1.1 (4), we obtain $\tau(s^*_k, \eta) = 1$ for $k = 1, \ldots, K-1$. As $\eta \in \mathcal{S}_1$, it achieves the maximum over the class $\mathcal{S}_1$, yielding (2). Hence, for any median scoring function $\bar{s}$, we have $\tau(s^*_k, \bar{s}) = 1$ for $k \in \{1, \ldots, K-1\}$, i.e. $\bar{s} \in \mathcal{S}^*_{k+1,k}$ for $k \in \{1, \ldots, K-1\}$, and thus $\bar{s} \in \mathcal{S}^*$.

Proof of Proposition 4.2.3

Recall that $\tau(s_1, s_2) = 1 - 2d_\tau(s_1, s_2)$, where $d_\tau(s_1, s_2)$ is given by:

$$\mathbb{P}\{(s_1(X)-s_1(X'))\cdot(s_2(X)-s_2(X')) < 0\}+\frac{1}{2}\mathbb{P}\{s_1(X) = s_1(X'),\ s_2(X) \neq s_2(X')\} + \frac{1}{2}\mathbb{P}\{s_1(X) \neq s_1(X'),\ s_2(X) = s_2(X')\}.$$

Observe first that, for all $s \in \mathcal{S}_0$, $\mathrm{AUC}_{\phi_1,\phi_2}(\mathrm{s})$ may be written as:

$$\mathbb{P}\{(s(X)-s(X'))\cdot(Y-Y') > 0\}/(2p(1-p))+\mathbb{P}\{s(X) = s(X'),\ Y \neq Y'\}/(4p(1-p)).$$

Notice also that, using Jensen's inequality, one easily obtain that the quantity $2p(1-p)|\mathrm{AUC}_{\phi_1,\phi_2}(\mathrm{s}_1) - \mathrm{AUC}_{\phi_1,\phi_2}(\mathrm{s}_2)|$ is bounded by the expectation of the random variable

$$\mathbb{I}\{(s_1(X)-s_1(X'))\cdot(s_2(X)-s_2(X')) < 0\}+\frac{1}{2}\mathbb{I}\{s_1(X) = s_1(X')\}\cdot\mathbb{I}\{s_2(X) \neq s_2(X')\}+ \frac{1}{2}\mathbb{I}\{s_1(X) \neq s_1(X')\}\cdot\mathbb{I}\{s_2(X) = s_2(X')\},$$

which is equal to $d_\tau(s_1, s_2) = (1 - \tau(s_1, s_2))/2$. This proves the assertion.

Proof of Proposition 4.2.4

Set $\Gamma_s = \{(x, x') \in \mathcal{X}^2 : \ (\zeta(x) - \zeta(x'))(s(x) - s(x')) < 0\}$. We have, for all real valued scoring functions $(s, s^*) \in \mathcal{S} \times \mathcal{S}^*_{1,2}$:

$$d_\tau(s, s^*) \leq \mathbb{P}\left\{(X, X') \in \Gamma_s\right\} + \frac{1}{2}\mathbb{P}\{s(X) = s(X')\}.$$

Recall also that

$$\begin{aligned} 2p(1-p)\left(\mathrm{AUC}^*_{\mathrm{f}_1,\mathrm{f}_2} - \mathrm{AUC}_{\mathrm{f}_1,\mathrm{f}_2}(\mathrm{s})\right) = \mathbb{E}\left[|\zeta(X) - \zeta(X')|\mathbb{I}\{(X, X') \in \Gamma_s\}\right] \\ + \mathbb{P}\{s(X) = s(X'),\ (Y, Y') = (-1, +1)\}, \end{aligned}$$

see Example 1 in [Clémençon *et al.*, 2008] for instance.

Observe that Hölder inequality combined with the noise condition shows that the quantity $\mathbb{E}\left[\mathbb{I}\{(X, X') \in \Gamma_s\}\right]$ is bounded by

$$\mathbb{E}\left[|\zeta(X) - \zeta(X')| \cdot \mathbb{I}\{(X, X') \in \Gamma_s\}\right]^{a/(1+a)} c^{1/(1+a)}.$$

In addition, we have

$$\begin{aligned} \mathbb{P}\{s(X) = s(X'),\ (Y, Y') = (-1, +1)\} \\ = \frac{1}{2}\mathbb{E}\left[\mathbb{I}\{s(X) = s(X')\} \cdot (\zeta(X) + \zeta(X') - 2\zeta(X)\zeta(X'))\right], \end{aligned}$$

and the upper bound can be easily seen as larger than $\mathbb{E}\left[\mathbb{I}\{s(X) = s(X')\} \cdot |\zeta(X) - \zeta(X')|\right]/2$. Therefore, using the same Hölder argument as above, we obtain that

$$\mathbb{P}\{s(X) = s(X')\} \leq \left(\mathbb{E}\left[|\zeta(X) - \zeta(X')| \cdot \mathbb{I}\{s(X) = s(X')\}\right)^{a/(1+a)} \times c^{1/(1+a)}.$$

Combining the bounds above, the concavity of $t \mapsto t^{a/(1+a)}$ permits to finish the proof.

Proof of Theorem 4.2.1

Let $(s_n^{(1)}, s_n^{(2)})$ be a sequence of real-valued scoring functions in $\mathcal{S}_1$ such that, as $n \to \infty$, $\mathrm{AUC}_{\mathrm{f}_1,\mathrm{f}_2}(\mathrm{s}_\mathrm{n}^{(1)}) \to \mathrm{AUC}^*_{\mathrm{f}_1,\mathrm{f}_2}$ and $\mathrm{AUC}_{\mathrm{f}_2,\mathrm{f}_3}(\mathrm{s}_\mathrm{n}^{(2)}) \to \mathrm{AUC}^*_{\mathrm{f}_2,\mathrm{f}_3}$. Here we consider the following consensus measure: $\forall s \in \mathcal{S}_1$,

$$\Delta_n(s) = d_\tau\left(s, s_n^{(1)}\right) + d_\tau\left(s, s_n^{(2)}\right).$$

Let $s^* \in \mathcal{S}_1 \cap \mathcal{S}^*$. Denote by $d_{\tau_{1,2}}$ the Kendall tau distance when $X \sim (p_1/(1 - p_3))F_1 + (p_2/(1-p_3))F_2$. Proposition 4.2.3, combined with the triangular inequality

applied to the pseudo-distance $d_{\tau_{1,2}}$, implies that

$$\begin{aligned}\mathrm{AUC}^*_{\mathrm{f}_1,\mathrm{f}_2} - \mathrm{AUC}_{\mathrm{f}_1,\mathrm{f}_2}(\bar{\mathrm{s}}_\mathrm{n}) &\leq \frac{d_{\tau_{1,2}}(s^*, \bar{s}_n)}{p_1p_2/(1-p_3)^2} \\ &\leq \frac{d_{\tau_{1,2}}(\bar{s}_n^{(1)}, \bar{s}_n) + d_{\tau_{1,2}}(s^*, \bar{s}_n^{(1)})}{p_1p_2/(1-p_3)^2} \\ &\leq \frac{d_{\tau_{1,2}}(s^*, \bar{s}_n^{(1)})}{p_1p_2/(1-p_3)^2} + \frac{d_\tau(\bar{s}_n^{(1)}, \bar{s}_n)}{p_1p_2}.\end{aligned}$$

The desired result finally follows from Proposition 4.2.4 combined with the AUC-consistency assumptions.

Proof of Proposition 4.3.2

Proof. Recall that $\tau_\mu(s_1, s_2) = 1 - 2d_{\tau_\mu}(s_1, s_2)$, where $d_{\tau_\mu}(s_1, s_2)$ is given by:

$$\begin{aligned}&\mathbb{P}_\mu\{(s_1(X) - s_1(X')) \cdot (s_2(X) - s_2(X')) < 0\} \\ &+\frac{1}{2}\mathbb{P}_\mu\{s_1(X) = s_1(X'),\ s_2(X) \neq s_2(X')\} + \frac{1}{2}\mathbb{P}_\mu\{s_1(X) \neq s_1(X'),\ s_2(X) = s_2(X')\}.\end{aligned}$$

Observe first that, for all $s \in \mathcal{S}_0$, $M_{1v23}(s)$ may be written as:

$$\begin{aligned}&\mathbb{P}\{s(X) < s(X'), Y = 1, Y' \in \{2,3\}\}/(p_1(1-p_1)) \\ &\qquad + \mathbb{P}\{s(X) = s(X'), Y = 1, Y' \in \{2,3\}\}/(2p_1(1-p_1)).\end{aligned}$$

Notice also that, using Jensen's inequality, one easily obtains that the quantity $2p_1(1-p_1)|M_{\mu,1v23}(s_1) - M_{\mu,1v23}(s_2)|$ is bounded by the expectation of the random variable

$$\begin{aligned}&\mathbb{I}\{(s_1(X) - s_1(X')) \cdot (s_2(X) - s_2(X')) < 0\} \\ &+ \frac{1}{2}\mathbb{I}\{s_1(X) = s_1(X'), s_2(X) \neq s_2(X')\} + \frac{1}{2}\mathbb{I}\{s_1(X) \neq s_1(X'), s_2(X) = s_2(X')\},\end{aligned}$$

which is equal to $d_{\tau_\mu}(s_1, s_2) = (1 - \tau_\mu(s_1, s_2))/2$. This proves the assertion. The proof is similar for $M_{\mu,12v3}(s)$. □

Proof of Proposition 4.3.3

Set $\Gamma_{s,i} = \{(x, x') \in \mathcal{X}^2 : (\eta_i(x) - \eta_i(x'))(s(x) - s(x')) < 0\}$. For all real valued scoring functions $(s, s^*) \in \mathcal{S} \times \mathcal{S}^*$:

$$d_{\tau_\mu}(s, s^*) \leq \mathbb{P}\left\{(X, X') \in \Gamma_s\right\} + \frac{1}{2}\mathbb{P}\{s(X) = s(X')\}.$$

Recall also that

$$\begin{aligned}2p_1(1-p_1)\left(M^*_{\mu,1v23} - M_{\mu,1v23}(s)\right) = \mathbb{E}\left[|\eta_1(X) - \eta_1(X')|\mathbb{I}\{(X, X') \in \Gamma_s\}\right] \\ + \mathbb{P}\{s(X) = s(X'),\ Y = 1, Y' \in \{2,3\}\}.\end{aligned}$$

Observe that according to Hölder inequality combined with the noise condition, the quantity $\mathbb{E}\left[\mathbb{I}\{(X,X')\in\Gamma_s\}\right]$ is bounded by

$$\mathbb{E}\left[|\eta_1(X)-\eta_1(X')|\mathbb{I}\{(X,X')\in\Gamma_s\}\right]^{a/(1+a)}c^{1/(1+a)}.$$

In addition, $\mathbb{P}\{s(X)=s(X'),\ Y=1,Y'\in\{2,3\}\}$ equals

$$\frac{1}{2}\mathbb{E}\left[\mathbb{I}\{s(X)=s(X')\}\cdot\eta_1(X)(1-\eta_1(X'))\right] + \frac{1}{2}\mathbb{E}\left[\mathbb{I}\{s(X)=s(X')\}\cdot\eta_1(X')(1-\eta_1(X))\right],$$

and the upper bound can be easily seen as larger than
$\mathbb{E}\left[\mathbb{I}\{s(X)=s(X')\}\cdot|\eta_1(X)-\eta_1(X')\right]/2$. Therefore, using the same Hölder argument as above, we obtain that

$$\mathbb{P}\{s(X)=s(X')\}\leq c^{\frac{1}{1+a}}\left(\mathbb{E}[|\eta_1(X)-\eta_1(X')|\cdot\mathbb{I}\{s(X)=s(X')\}]\right)^{\frac{a}{1+a}}$$

Combining the bounds above, the concavity of $t\mapsto t^{a/(1+a)}$enables to finish the proof. The proof is similar for $\mathrm{AUC}_{\mu,12\mathrm{v}3}(\mathrm{s})$.

Proof of Theorem 4.3.4

Let $(s_n^{(1)},s_n^{(2)})$ be a sequence of real-valued scoring functions in $\mathcal{S}_1$ such that, as $n\to\infty$, $M_{1v23}(s_n^{(1)})\to M^*_{1v23}$ and $M_{12v3}(s_n^{(2)})\to M^*_{12v3}$. Here we consider the following consensus measure: $\forall s\in\mathcal{S}_1$,

$$\Delta_n(s)=d_{\tau_\mu}\left(s,s_n^{(1)}\right)+d_{\tau_\mu}\left(s,s_n^{(2)}\right).$$

Let $s^*\in\mathcal{S}_1\cap\mathcal{S}^*$. It is easy to show that
$\left(\mathrm{AUC}^*_{\mathrm{F}_1,\mathrm{F}_2}-\mathrm{AUC}_{\mathrm{F}_1,\mathrm{F}_2}(\mathrm{s})\right)\leq\frac{1}{1-p_3}\left(M^*_{\mu,1v2}-M_{\mu,1v2}(s)\right)$ and using the MLR assumption that this is bounded by $\frac{1-p_1}{p_1(1-p_3)}(M^*_{1v23}-M_{1v23}(s))$. Using the theorem 1.2.12, we get

$$\mathrm{VUS}^*-\mathrm{VUS}(\bar{\mathrm{s}}_\mathrm{n})\leq\frac{1-\mathrm{p}_1}{\mathrm{p}_1(1-\mathrm{p}_3)}\left(\mathrm{M}^*_{1\mathrm{v}23}-\mathrm{M}_{1\mathrm{v}23}(\bar{\mathrm{s}}_\mathrm{n})\right) + \frac{1-p_3}{p_3(1-p_1)}\left(M^*_{12v3}-M_{12v3}(\bar{s}_n)\right).$$

Using the Proposition 4.3.2, the right hand side can be upper bounded by $\frac{d_{\tau_\mu}(s^*,\bar{s}_n)}{p_1^2(1-p_3)}+\frac{d_{\tau_\mu}(s^*,\bar{s}_n)}{p_3^2(1-p_1)}$. Let $C=\max\{\frac{1}{p_1^2(1-p_3)},\frac{1}{p_3^2(1-p_1)}\}$. Combining the triangular inequality and the definition of $\bar{s}_n$, we obtain

$$\begin{aligned}\mathrm{VUS}^* - \mathrm{VUS}(\bar{\mathrm{s}}_\mathrm{n}) &\leq \frac{\mathrm{d}_{\tau_\mu}(\mathrm{s}^*,\mathrm{s}^{(1)})+\mathrm{d}_{\tau_\mu}(\mathrm{s}^{(1)},\bar{\mathrm{s}}_\mathrm{n})}{\mathrm{p}_1^2(1-\mathrm{p}_3)}+\frac{\mathrm{d}_{\tau_\mu}(\mathrm{s}^*,\mathrm{s}^{(2)})+\mathrm{d}_{\tau_\mu}(\mathrm{s}^{(2)},\bar{\mathrm{s}}_\mathrm{n})}{\mathrm{p}_3^2(1-\mathrm{p}_1)}\\ &\leq C\left(d_{\tau_\mu}(s^*,s^{(1)})+d_{\tau_\mu}(s^{(1)},\bar{s}_n)\right)+C\left(d_{\tau_\mu}(s^*,s^{(2)})+d_{\tau_\mu}(s^{(2)},\bar{s}_n)\right)\\ &\leq 2C\left(d_{\tau_\mu}(s^*,s^{(1)})+d_{\tau_\mu}(s^*,s^{(2)})\right).\end{aligned}$$

The desired result finally follows from Proposition 4.3.3 combined with the M-consistency assumptions.

Chapter 5

TreeRank Tournament

The goal of this chapter is to build piecewise constant scoring functions for multipartite ranking problem using an approximation scheme of the optimal ROC surface. This strategy is at the origin of the algorithm TreeRank (see [Clémençon & Vayatis, 2009b], [Baskiotis *et al.*, 2010] and [Clémençon *et al.*, 2011a]) for the bipartite ranking and we adapt it for the multipartite ranking problem.

A particular class of learning algorithms are considered taking the form of decision trees in the spirit of CART, see [Breiman *et al.*, 1984]. These tree-based procedures recursively build a partition of the input space $\mathcal{X}$. In classification and regression, the predicted labels depend only on the subregion containing the observations so the splitting rules of the decision tree are based on local learning. However, the ranking problem is a global learning task and the notion of ordering would rather involve comparing sub-regions to each other. Several adaptations of decision trees in the context of ranking have been proposed (see [Ferri *et al.*, 2002], [Provost & Domingos, 2003], [Xia *et al.*, 2006]) and they mainly rely on changing the splitting rules for practical matter.

For the output of the TreeRank procedure, the ordering is tree-structured i.e. ranks are read from left to right at the bottom of the resulting tree. This simple top-down algorithm, is interpreted as a statistical counterpart of an adaptive and recursive piecewise linear approximation of the ROC curve, similarly to the finite element methods. In that case, the problem of recovering the optimal ROC curve and the problem of adaptively building a scoring function from training data with a ROC curve close to the approximate version of the optimal one can be addressed simultaneously. Moreover, the splitting rule can be rewritten as a weighted classification problem.

The main difficulty of extending the TreeRank procedure relies on the fact that the optimal splitting rule can not be interpreted as a weighted classification problem. To overcome this issue, a tournament between binary classifiers learned from a pair of labels is organized. Finally, the classifier that maximizes the VUS is chosen as a splitting rule. Although this choice may be sub-optimal, this procedure provides an approximation of the optimal ROC surface with the same rate as the TreeRank procedure for the $L_\infty-$norm and the L_1-norm. Moreover, this simple strategy can be used for any number of classes and several decompositions of the multipartite problem can be considered.

The chapter is structured as follows. In section 5.1, we present the main properties of the piecewise constant scoring functions and the behavior of the associated ROC surface. In section 5.2, we present the approximation scheme of the optimal ROC surface and state its consistency. In section 5.3, we describe the practical implementation of the TreeRank Tournament procedure and we state theoretical results. All proofs are postponed to the section 5.5.

5.1 Background and Preliminaries

It is the purpose of this section to recall crucial notions inherent to the formulation of the multipartite ranking issue and to performance evaluation in this context.

5.1.1 Bipartite Ranking and the TreeRank Algorithm

In [Clémençon & Vayatis, 2009b] (see also [Clémençon *et al.*, 2011a]), a bipartite ranking algorithm optimizing directly the ROC curve in a recursive manner, called TreeRank, has been proposed and thoroughly studied. It produces an oriented partition of the feature space $\mathcal{X}$ (defining thus a ranking, for which elements of a same cell being viewed as ties). The process is described by a left-to-right oriented binary tree structure, termed *ranking tree*, with fixed maximum depth $J \geq 0$. At depth $j \leq J$, there are 2^j nodes, indexed by (j,k) with $0 \leq k < 2^j$. The root node represents the whole feature space $\mathcal{C}_{0,0} = \mathcal{X}$ and each *internal node* (j,k) with $j < J$ and $k \in \{0, \ldots, 2^j - 1\}$ corresponds to a subset $\mathcal{C}_{j,k} \subset \mathcal{X}$, whose right and left siblings respectively depict disjoint subsets $\mathcal{C}_{j+1,2k}$ and $\mathcal{C}_{j+1,2k+1}$ such that $\mathcal{C}_{j,k} = \mathcal{C}_{j+1,2k} \cup \mathcal{C}_{j+1,2k+1}$. At the root, one starts with a constant scoring function $s_1(x) = \mathbb{I}\{x \in \mathcal{C}_{0,0}\} \equiv 1$ and after $m = 2^j + k$ iterations, $0 \leq k < 2^j$, the current scoring function is $s_m(x) = \sum_{l=0}^{2k-1}(m-l)\cdot\mathbb{I}\{x \in \mathcal{C}_{j+1,l}\} + \sum_{l=k}^{2^j-1}(m-k-l)\cdot\mathbb{I}\{x \in \mathcal{C}_{j,l}\}$ and the cell $\mathcal{C}_{j,k}$ is split in order to form an updated version of the scoring function, $s_{m+1}(x) = \sum_{l=0}^{2k}(m-l)\cdot\mathbb{I}\{x \in \mathcal{C}_{j+1,l}\} + \sum_{l=k+1}^{2^j-1}(m-k-l)\cdot\mathbb{I}\{x \in \mathcal{C}_{j,l}\}$ namely, with maximum (empirical) AUC. Therefore, it happens that this problem boils down to solve a cost-sensitive binary classification problem on the set $\mathcal{C}_{j,k}$, see subsection 3.3 in [Clémençon *et al.*, 2011a] for further details. Indeed, one may write the AUC increment as

$$\mathrm{AUC}_{1,2}(\mathrm{s_{m+1}}) - \mathrm{AUC}_{1,2}(\mathrm{s_m}) = \frac{1}{2}\mathrm{F_1}(\mathcal{C}_{\mathrm{j,k}})\mathrm{F_2}(\mathcal{C}_{\mathrm{j,k}})(1 - \Lambda_{1,2}(\mathcal{C}_{\mathrm{j+1,2k}} \mid \mathcal{C}_{\mathrm{j,k}})),$$

where $\Lambda_{1,2}(\mathcal{C}_{j+1,2k} \mid \mathcal{C}_{j,k}) \overset{def}{=} F_2(\mathcal{C}_{j,k} \setminus \mathcal{C}_{j+1,2k})/F_2(\mathcal{C}_{j,k}) + F_1(\mathcal{C}_{j+1,2k})/F_1(\mathcal{C}_{j,k})$. Setting $p = F_2(\mathcal{C}_{j,k})/(F_1(\mathcal{C}_{j,k}) + F_2(\mathcal{C}_{j,k}))$, the crucial point of the TreeRank approach is that the quantity $2p(1-p)\Lambda_{1,2}(\mathcal{C}_{j+1,2k} \mid \mathcal{C}_{j,k})$ can be seen as the cost-sensitive error of a classifier on $\mathcal{C}_{j,k}$ predicting label 2 on $\mathcal{C}_{j+1,2k}$ and label 1 on $\mathcal{C}_{j,k} \setminus \mathcal{C}_{j+1,2k}$ with cost p (respectively, $1-p$) assigned to the error consisting in predicting label 2 given $Y = 1$ (resp., label 1 given $Y = 2$), balancing thus the two types of error. Hence, at

each iteration of the ranking tree growing stage, the TreeRank algorithm calls a *cost-sensitive* binary classification algorithm, termed LeafRank, in order to solve a statistical version of the problem above (replacing the theoretical probabilities involved by their empirical counterparts) and split $\mathcal{C}_{j,k}$ into $\mathcal{C}_{j+1,2k}$ and $\mathcal{C}_{j+1,2k+1}$. As described at length in [Clémençon *et al.*, 2011a], one may use cost-sensitive versions of celebrated binary classification algorithms such as CART or SVM for instance as LeafRank procedure, the performance depending on their ability to capture the geometry of the level sets of the likelihood ratio $dF_2/dF_1(x)$. In general, the growing stage is followed by a pruning procedure, where children of a same parent node are recursively merged in order to produce a ranking subtree that maximizes an estimate of the AUC criterion, based on cross-validation usually (*cf* section 4 in [Clémençon *et al.*, 2011a]). Under adequate assumptions, consistency results and rate bounds for the TreeRank approach (in the sup norm sense and for the AUC deficit both at the same time) are established in [Clémençon & Vayatis, 2009b] and [Clémençon *et al.*, 2011a], an extensive experimental study can be found in [Clémençon *et al.*, 2012].

5.1.2 Multipartite Ranking Algorithms

In contrast to the bipartite situation (see [Clémençon & Vayatis, 2010], [Clémençon & Vayatis, 2009b]), no algorithm optimizing the ROC surface directly and producing a scoring function $\widehat{s}_n$ for which $d_\infty(\widehat{s}_n, s^*) \to 0$ in probability has been documented in the literature. Beyond theoretical results guaranteeing the validity of empirical maximization of the VUS criterion (see [Rajaram & Agarwal, 2005]), most methods proposed rely on the optimization of an alternative (pairwise) criterion ([Freund *et al.*, 2003] and [Pahikkala *et al.*, 2007]), or on the decomposition of the original multipartite problem into bipartite subproblems combined with a final aggregation/consensus stage ([Hüllermeier *et al.*, 2008] and [Clémençon *et al.*, 2013b]) or still on plug-in approaches based on ordinal regression ([Waegeman *et al.*, 2008c]). In addition, it is far from straightforward to extend the TreeRank algorithm recalled above because, when $K \geq 3$, as a straightforward computation based on Eq. 1.2.9 may show, the splitting step cannot be interpreted as a learning problem which can be solved by means of off-the-shelf techniques, unlike the bipartite case. Indeed, taking $s(x) = \mathbb{I}\{x \in \mathcal{C}\}$ for some measurable set $\mathcal{C} \subset \mathcal{X}$, we have

$$\mathrm{VUS(s)} = \mathrm{F}_3(\mathcal{C})(1-\mathrm{F}_1(\mathcal{C}))/2 + (1-\mathrm{F}_1(\mathcal{C}))(1-\mathrm{F}_2(\mathcal{C}))(1-\mathrm{F}_3(\mathcal{C}))/6 + F_1(\mathcal{C})F_2(\mathcal{C})F_3(\mathcal{C})/6. \quad (5.1)$$

It is the goal of this paper to propose an alternative, letting splitting rule candidates, corresponding to solutions of different bipartite subproblems, compete for VUS maximization in a tournament.

5.1.3 Further Notations and Preliminaries

Let $\mathcal{P} = (\mathcal{C}_j)_{1\leq j\leq N}$ be an *ordered partition* of the input space $\mathcal{X}$ counting $N \geq 1$ cells. The adjective *ordered* means here that, for any $1 \leq i \leq j \leq N$, instances lying in $\mathcal{C}_i$ are expected to have higher labels than those in $\mathcal{C}_j$, in a way that $\mathcal{P}$ is related to the scoring function

$$S_{\mathcal{P}}(x) = \sum_{i=1}^{N} (N - i + 1) \cdot \mathbb{I}\{x \in \mathcal{C}_i\}.$$

We point out that in the tripartite case, $S_{\mathcal{P}}$'s ROC surface is piecewise planar with N^2 pieces. More precisely, it is the polytope that connects the points

$$\left(1 - \sum_{l=1}^{i} F_1(\mathcal{C}_i),\ \sum_{l=j+1}^{i} F_2(\mathcal{C}_l),\ \sum_{l=1}^{j} F_3(\mathcal{C}_l)\right),$$

where $0 \leq j \leq i \leq N$, with the convention that empty summation equals zero. In order to provide a closed analytical form for the latter, set $\alpha_j = 1 - F_1(\cup_{l=1}^{j} \mathcal{C}_l)$ and $\gamma_j = F_3(\cup_{l=1}^{j} \mathcal{C}_l)$ for $j = 1, \ldots, N$ and $1 - \alpha_0 = 0 = \gamma_0$ by convention. Set also $\phi(\alpha, \alpha', \alpha'') = \frac{\alpha - \alpha'}{\alpha'' - \alpha'} \mathbb{I}\{\alpha \in [\alpha'; \alpha'']\}$ for all $\alpha' \leq \alpha \leq \alpha''$ and consider the *hat functions* defined by $\phi_i(\alpha) = \phi(\alpha, \alpha_{i-1}, \alpha_i) - \phi(\alpha, \alpha_i, \alpha_{i+1})$ and $\varphi_j(\gamma) = \phi(\gamma, \gamma_{j-1}, \gamma_j) - \phi(\gamma, \gamma_j, \gamma_{j+1})$, as well as the tensorial products $\Phi_{i,j}(\alpha, \gamma) = \phi_i(\alpha)\varphi_j(\gamma)$ for $1 \leq i,\ j \leq N$, which are the basis functions used in the *Finite Element Method* to approximate real valued functions defined on $[0, 1]^2$. Equipped with these notations, the ROC surface of $S_{\mathcal{P}}$ can be written as

$$\mathrm{ROC}_{\mathrm{S}_{\mathcal{P}}}(\alpha, \gamma) = \sum_{1\leq \mathrm{j}\leq \mathrm{i}\leq \mathrm{N}}^{\mathrm{N}} \mathrm{F}_2\left(\bigcup_{\mathrm{l}=1+\mathrm{j}}^{\mathrm{i}} \mathcal{C}_{\mathrm{l}}\right) \Phi_{\mathrm{i,j}}(\alpha, \gamma). \tag{5.2}$$

5.2 Adaptive Piecewise Planar Approximation of ROC*

This section is dedicated to the analysis of an adaptive approximation scheme of the optimal ROC surface, which outputs a piecewise planar approximate of ROC* that is itself the ROC surface of a piecewise constant scoring function. In order to describe it at length, further notations are required.

5.2.1 An Implicit Tree-Structured Recursive Interpolation Scheme

Here, we describe a recursive approximation scheme to build a piecewise constant scoring function $S^*_{\mathcal{P}^*}$ whose ROC surface can be viewed as a piecewise planar interpolant of ROC*, corresponding to a mesh grid adaptively chosen. As shall be seen below, the related oriented partition $\mathcal{P}^*$ can be represented by means of a left-to-right oriented binary tree structure $\{\mathcal{C}^*_{j,l} :\ j \leq J,\ l = 0, \ldots, 2^j - 1\}$ and its cells coincide with certain bilevel sets of the regression function $\eta(x)$. In addition,

as shall be seen below, the distance (in sup-norm) between $\mathrm{ROC}_{\mathrm{S}^*_{\mathcal{P}^*}}$ and ROC^* can be bounded as a function of the number of iterations (*i.e.* of the number of cells of $\mathcal{P}^*$) under the following assumptions.

Assumption 3. *The class distributions F_1, F_2 and F_3 are equivalent and the likelihood ratios $\Phi_{2,1}, \Phi_{3,1}, \Phi_{3,2}$ are bounded.*

Assumption 4. *The distribution of $\eta(X)$ is absolutely continuous with respect to Lebesgue measure. Let $F_k^*(x)$ and $F_k^*(dx) = f_k^*(x)dx$ be the conditional cdf and df of $\eta(X)$ given $Y = k$, with $1 \leq k \leq 3$.*

In particular, these hypotheses guarantee that the optimal ROC surface exhibits a minimum amount of smoothness, as stated in the proposition below.

Proposition 5.2.1. *Under* **Assumptions 1-3**, *the mapping $(\alpha, \gamma) \in [0,1]^2 \mapsto \mathrm{ROC}^*(\alpha, \gamma)$ is differentiable. On the set $\mathcal{I}^* = \{(\alpha,\ \gamma) \in [0,1]^2 :\ F_2^* \circ F_3^{*-1}(1-\gamma) \geq F_2^* \circ F_1^{*-1}(\alpha)\}$, the first partial derivatives of ROC^* are given by:*

$$\frac{\partial}{\partial\alpha}\mathrm{ROC}^*(\alpha,\gamma) = -\frac{\mathrm{f}_2^*}{\mathrm{f}_1^*}(\mathrm{F}_1^{*-1}(\alpha)),\ \ \frac{\partial}{\partial\gamma}\mathrm{ROC}^*(\alpha,\gamma) = -\frac{\mathrm{f}_2^*}{\mathrm{f}_3^*}(\mathrm{F}_1^{*-1}(1-\gamma)).$$

They are equal to zero on the complementary set $[0,1]^2 \setminus \mathcal{I}^*$.

The subsequent analysis actually actually requires that a slightly stronger smoothness assumption holds true.

Assumption 5. *The mapping ROC^* is twice differentiable with bounded second derivatives given by: $\forall(\alpha,\gamma) \in \mathcal{I}^*$,*

$$\frac{\partial^2}{\partial\alpha^2}\mathrm{ROC}^*(\alpha,\gamma) = -\frac{\mathrm{f}_2'^*\mathrm{f}_1^* - \mathrm{f}_2^*\mathrm{f}_1'^*}{\mathrm{f}_1^{*3}}(\mathrm{F}_1^{*-1}(\alpha)),\ \ \frac{\partial^2}{\partial\gamma^2}\mathrm{ROC}^*(\alpha,\gamma) = \frac{\mathrm{f}_2'^*\mathrm{f}_3^* - \mathrm{f}_2^*\mathrm{f}_3'^*}{\mathrm{f}_3^{*3}}(\mathrm{F}_3^{*-1}(1-\gamma)).$$

Initialization. We set $\mathcal{C}_{0,0}^* = \mathcal{X}$, $s_1^*(x) \equiv 1$ and $1 = \alpha_{0,0}^* = 1 - \alpha_{0,1}^* = 1 - \gamma_{0,0}^* = \gamma_{0,1}^* = 1 - \beta_{0,0}^* = \beta_{0,1}^*$. Observe that $F_1(\mathcal{C}_{0,0}^*) = \alpha_{0,0}^* - \alpha_{0,1}^*$, $F_2(\mathcal{C}_{0,0}^*) = \beta_{0,1}^* - \beta_{0,0}^*$ and $F_3(\mathcal{C}_{0,0}^*) = \gamma_{0,1}^* - \gamma_{0,0}^*$. In the $\alpha\gamma\beta$ system of coordinates, the initial approximant of the surface ROC^* is the planar piece connecting $(1,0,0) = (\alpha_{0,0}^*, \gamma_{0,0}^*, \beta_{0,0}^*)$, $(0,1,0) = (\alpha_{0,0}^*, \gamma_{0,1}^*, \beta_{0,0}^*)$ and $(0,0,1) = (\alpha_{0,1}^*, \gamma_{0,0}^*, \beta_{0,1}^*)$. It is the surface $\{(\alpha,\ \gamma, \widetilde{\mathrm{ROC}}_1^*(\alpha,\gamma)) :\ (\alpha,\gamma) \in [0,1]^2\}$ with $\widetilde{\mathrm{ROC}}_1^*(\alpha,\gamma) = 1-\alpha-\gamma$.

Iterations. For $j = 0,\ \ldots,\ J-1$ and for $k = 0,\ \ldots,\ 2^j - 1$:

• **Updates.** Set $\alpha_{j+1,2k}^* = \alpha_{j,k}^*$, $\alpha_{j+1,2k+2}^* = \alpha_{j,k+1}^*$, $\gamma_{j+1,2k}^* = \gamma_{j,k}^*$ and $\gamma_{j+1,2k+2}^* = \gamma_{j,k+1}^*$, $\beta_{j+1,2k}^* = \beta_{j,k}^*$ and $\beta_{j+1,2k+2}^* = \beta_{j,k+1}^*$.

• **Breakpoint candidates.** Considering the curve formed by the intersection between the current approximant of ROC* and the facet "$\gamma = 0$", define the point of coordinate

$$\alpha^{(1)}_{j+1,2k+1} = \mathrm{ROC}'^{*-1}_{1,2}\left(\frac{\beta^*_{\mathrm{j,k+1}} - \beta^*_{\mathrm{j,k}}}{\alpha^*_{\mathrm{j,k}} - \alpha^*_{\mathrm{j,k+1}}}\right)$$

on the α axis. This corresponds to the largest increase of the area under the curve when adding a breakpoint between $\alpha^*_{j,k}$ and $\alpha^*_{j,k+1}$, see Proposition 11 in [Clémençon & Vayatis, 2009b]. Incidentally, the resulting broken line is also optimal in the sup norm sense. Observe also that $\alpha^{(1)}_{j+1,2k+1} = \alpha^*_{j+1,2k} - F_1(\mathcal{C}^{(1)}_{j+1,2k})$, where $\mathcal{C}^{(1)}_{j+1,2k} = \arg\max_{\mathcal{C}\subset\mathcal{C}_{j,k}} \Lambda_{1,2}(\mathcal{C} \mid \mathcal{C}_{j,k})$. In addition, we have $\mathcal{C}^{(1)}_{j+1,2k} = \{x \in \mathcal{X} : F^{-1}_{\Phi_{1,2},1}(\alpha_{j+1,2k+1}) < \Phi_{1,2}(x) \leq F^{-1}_{\Phi_{1,2},1}(\alpha_{j+1,2k})\}$, where $F^{-1}_{\Phi_{1,2},1}(\alpha)$ denotes the quantile of order α of $\Phi_{1,2}(X)$'s conditional distribution given $Y = 1$. We also set $\beta^{(1)}_{j+1,2k+1} = \beta^*_{j+1,2k} + F_2(\mathcal{C}^{(1)}_{j+1,2k}) = \mathrm{ROC}^*_{1,2}(1 - \alpha^{(1)}_{\mathrm{j+1,2k+1}})$ and $\gamma^{(1)}_{j+1,2k+1} = \gamma^*_{j+1,2k} + F_3(\mathcal{C}^{(1)}_{j+1,2k}) = \mathrm{ROC}^*_{1,3}(1 - \alpha^{(1)}_{\mathrm{j+1,2k+1}})$.

In the same fashion, considering the curve formed by the intersection between the current approximate of ROC* and the facet "$\alpha = 0$", define the point of coordinate

$$\gamma^{(2)}_{j+1,2k+1} = \mathrm{ROC}'^{*-1}_{2,3}\left(\frac{\gamma^*_{\mathrm{j,k+1}} - \gamma^*_{\mathrm{j,k}}}{\beta^*_{\mathrm{j,k+1}} - \beta^*_{\mathrm{j,k}}}\right)$$

on the γ axis. This corresponds to the largest increase of the area under the curve when adding a breakpoint between $\gamma^*_{j,k}$ and $\gamma^*_{j,k+1}$. We have $\gamma^{(2)}_{j+1,2k+1} = \gamma^*_{j+1,2k} + F_3(\mathcal{C}^{(2)}_{j+1,2k})$, where $\mathcal{C}^{(2)}_{j+1,2k} = \arg\max_{\mathcal{C}\subset\mathcal{C}_{j,k}} \Lambda_{2,3}(\mathcal{C} \mid \mathcal{C}_{j,k})$. In addition, we have $\mathcal{C}^{(2)}_{j+1,2k} = \{x \in \mathcal{X} : F^{-1}_{\Phi_{2,3},3}(\gamma_{j+1,2k+1}) < \Phi_{2,3}(x) \leq F^{-1}_{\Phi_{2,3},1}(\gamma_{j+1,2k})\}$, where $F^{-1}_{\Phi_{2,3},3}(\gamma)$ denotes the quantile of order γ of $\Phi_{2,3}(X)$'s conditional distribution given $Y = 3$. We also set $\alpha^{(2)}_{j+1,2k+1} = \alpha^*_{j+1,2k} - F_1(\mathcal{C}^{(2)}_{j+1,2k}) = 1 - \mathrm{ROC}^*_{3,1}(\gamma^{(2)}_{\mathrm{j+1,2k+1}})$ and $\beta^{(2)}_{j+1,2k+1} = \beta^*_{j+1,2k} + F_2(\mathcal{C}^{(2)}_{j+1,2k}) = \mathrm{ROC}^*_{3,2}(\gamma^{(2)}_{\mathrm{j+1,2k+1}})$.

• **Tournament.** For $l \in \{1,\ 2\}$, compute the quantity

$$\begin{aligned}
\mathrm{VUS}_{\mathcal{C}^*_{\mathrm{j,k}}}(\mathcal{C}^{(\mathrm{l})}_{\mathrm{j+1,2k}}) &= \mathrm{F}_3(\mathcal{C}^{(\mathrm{l})}_{\mathrm{j+1,2k}})(\mathrm{F}_1(\mathcal{C}^*_{\mathrm{j,k}}) - \mathrm{F}_1(\mathcal{C}^{(\mathrm{l})}_{\mathrm{j+1,2k}}))/2 \\
&\quad + F_1(\mathcal{C}^{(l)}_{j+1,2k})F_2(\mathcal{C}^{(l)}_{j+1,2k})F_3(\mathcal{C}^{(l)}_{j+1,2k})/6 \\
&\quad + (F_1(\mathcal{C}^*_{j,k}) - F_1(\mathcal{C}^{(l)}_{j+1,2k}))(F_2(\mathcal{C}^*_{j,k}) - F_2(\mathcal{C}^{(l)}_{j+1,2k}))(F_3(\mathcal{C}^*_{j,k}) - F_3(\mathcal{C}^{(l)}_{j+1,2k}))/6 \\
&= (\gamma^{(l)}_{j+1,2k+1} - \gamma^*_{j+1,2k})(\alpha^{(l)}_{j+1,2k+1} - \alpha^*_{j+1,2k+2})/2 \\
&\quad + (\alpha^*_{j+1,2k} - \alpha^{(l)}_{j+1,2k+1})(\beta^{(l)}_{j+1,2k+1} - \beta^*_{j+1,2k})(\gamma^{(l)}_{j+1,2k+1} - \gamma^*_{j+1,2k})/6 \\
&\quad + (-\alpha^*_{j+1,2k+2} + \alpha^{(l)}_{j+1,2k+1})(-\beta^{(l)}_{j+1,2k+1} + \beta^*_{j+1,2k+2})(-\gamma^{(l)}_{j+1,2k+1} + \gamma^*_{j+1,2k+2})/6.
\end{aligned}$$

Then, determine $l^* = \arg\max_{l=1,\,2} \mathrm{VUS}_{\mathcal{C}^*_{j,k}}(\mathcal{C}^{(l)}_{j+1,2k})$ and set $\mathcal{C}^*_{j+1,2k} = \mathcal{C}^{(l^*)}_{j+1,2k}$ and $\mathcal{C}^*_{j+1,2k+1} = \mathcal{C}^*{}_{j,k} \setminus \mathcal{C}^{(l^*)}_{j+1,2k}$. This step is illustrated by Fig. 1 in the Supplementary Material. In addition, define $\alpha^*_{j+1,2k+1} = \alpha^*_{j+1,2k} - F_1(\mathcal{C}^*_{j+1,2k})$, $\beta^*_{j+1,2k+1} = \beta^*_{j+1,2k} + F_2(\mathcal{C}^*_{j+1,2k})$ and $\gamma^*_{j+1,2k+1} = \gamma^*_{j+1,2k} + F_3(\mathcal{C}^*_{j+1,2k})$.

Output. Compute the approximate given by: $\forall (\alpha, \gamma) \in [0,1]^2$,

$$\widetilde{\mathrm{ROC}}^*_J(\alpha, \gamma) = \sum_{1 \leq l \leq i \leq 2^J - 1} \left(\beta^*_{J,i} - \beta^*_{J,l}\right) \Phi^*_{i,l}(\alpha, \gamma),$$

where, for $1 \leq i,\ l \leq 2^J - 1$, we have set $\Phi_{i,l}(\alpha, \gamma) = \phi^*_i(\alpha)\varphi^*_l(\gamma)$ with $\phi^*_i(\alpha) = \phi(\alpha, \alpha^*_{J,i-1}, \alpha^*_{J,i}) - \phi(\alpha, \alpha^*_{J,i}, \alpha^*_{J,i+1})$ and $\varphi_l(\gamma) = \phi(\gamma, \gamma^*_{J,l-1}, \gamma^*_{J,l}) - \phi(\gamma, \gamma^*_{J,l}, \gamma^*_{J,l+1})$. Observe that it is the ROC surface of the scoring function:

$$s^*_{2^J}(x) = \sum_{l=0}^{2^J - 1} (2^J - l) \cdot \mathbb{I}\{x \in \mathcal{C}^*_{J,l}\}.$$

Indeed, we have: $\widetilde{\mathrm{ROC}}^*_J(\alpha, \gamma) = \mathrm{ROC}_{s^*_J}(\alpha, \gamma)$ for all $(\alpha, \gamma) \in [0,1]^2$.

It is noteworthy that the interpolant of the optimal ROC surface produced by the algorithm above is itself a (concave) ROC surface. Obviously, this is not the case in general, *cf* Eq. (5.2) above. This strikingly differs from the bipartite case, where any interpolant of the optimal ROC curve is the ROC curve of a piecewise constant scoring function, constant on certain bilevel sets of the likelihood ratio related to the class distributions, see subsection 3.1 in [Clémençon & Vayatis, 2010].

The following result provides guarantees for the approximation scheme described above.

Proposition 5.2.2. *Under* **Assumptions 1, 3, 4**, *there exists a constant* $C < +\infty$ *such that:*

$$\forall J \geq 1, \quad d_\infty(s^*, s^*_{2^J}) \leq C \times 2^{-2J}.$$

Now, the TreeRank Tournament algorithm can be clearly viewed as a statistical version of the interpolation scheme above. It will mimic it well, provided that each tournament yields a splitting rule closed to that based on the true VUS increment. This is the key to establish the rate bounds displayed in the next section.

5.3 Analysis of TreeRank Tournament

5.3.1 The TreeRank Tournament algorithm

We now describe the algorithm we propose to solve the multipartite ranking problem. We place ourselves in the tripartite case for notational simplicity, but extension to the general multipartite setting is straightforward, *cf* Chapter 1. The algorithm is implemented from a training dataset $\mathcal{D}_n = \{(X_1,\ Y_1),\ \ldots,\ (X_n,\ Y_n)\}$ and recursively calls a *cost-sensitive* binary classification algorithm $\mathcal{L}$ (*e.g.* SVM, CART,

Random Forest, k-NN), referred to as LeafRank. When ran on a set $\mathcal{C} \subset \mathcal{X}$, we denote by $\mathcal{L}(\mathcal{C})$ the collection of subsets of $\mathcal{C}$ over which algorithm $\mathcal{L}$ performs optimization. For $1 \leq k \leq 3$, set $n_k = \sum_{i=1}^{n} \mathbb{I}\{Y_i = k\}$ and define, for any measurable set $\mathcal{C} \subset \mathcal{X}$, $\widehat{F}_k(\mathcal{C}) = (1/n_k) \sum_{i=1}^{n} \mathbb{I}\{X_i \in \mathcal{C},\ Y_i = k\}$, and, for any measurable subset $\mathcal{C}' \subset \mathcal{C}$ with $1 \leq k < l \leq 3$,

$$\widehat{\Lambda}_{k,l}(\mathcal{C}' \mid \mathcal{C}) = \widehat{F}_l(\mathcal{C}')/\widehat{F}_l(\mathcal{C}) + \widehat{F}_k(\mathcal{C} \setminus \mathcal{C}')/\widehat{F}_k(\mathcal{C}).$$

As already pointed out in subsection 5.1.1, the quantity above can be seen as proportional to the empirical *cost-sensitive* error of a binary classifier on the restricted input space $\mathcal{C}$ which predicts label l on $\mathcal{C}'$ and label k on $\mathcal{C} \setminus \mathcal{C}'$ with cost $\widehat{F}_l(\mathcal{C})/(\widehat{F}_k(\mathcal{C}) + \widehat{F}_l(\mathcal{C}))$ (respectively, $\widehat{F}_k(\mathcal{C})/(\widehat{F}_k(\mathcal{C}) + \widehat{F}_l(\mathcal{C}))$) assigned to the error consisting in predicting label l while the true label is k (resp., label k, while the true label is l), based on the data of the original sample $\mathcal{D}_n$ lying in the set $\mathcal{C}$ with label k or l. We also introduce the quantity:

$$\begin{aligned}\widehat{\mathrm{VUS}}_{\mathcal{C}}(\mathcal{C}') = \widehat{F}_3(\mathcal{C}')(\widehat{F}_1(\mathcal{C}) - \widehat{F}_1(\mathcal{C}'))/2 + \widehat{F}_1(\mathcal{C}')\widehat{F}_2(\mathcal{C}')\widehat{F}_3(\mathcal{C}')/6 \\ + (\widehat{F}_1(\mathcal{C}) - \widehat{F}_1(\mathcal{C}'))(\widehat{F}_2(\mathcal{C}) - \widehat{F}_2(\mathcal{C}'))(\widehat{F}_3(\mathcal{C}) - \widehat{F}_3(\mathcal{C}'))/6,\end{aligned}$$

which corresponds to the empirical VUS increase resulting from splitting the cell $\mathcal{C}$ into left and right siblings $\mathcal{C}'$ and $\mathcal{C} \setminus \mathcal{C}'$, *cf* Eq. (5.1).

A straightforward variant of the algorithm above could consist in running additionally the LeafRank algorithm for local cost-sensitive classification problems related to the pair of labels $(1, 3)$ and thus enlarging the set of competitors ("extended tournament"). This would however increase the amount of computations performed. In addition, just like for TreeRank algorithm in the bipartite context (see [Clémençon *et al.*, 2011a]), the ranking tree growing procedure described above can be followed by a PRUNING STAGE in order to maximize a (cross-validation based) estimate of the VUS criterion. Model selection analysis is however beyond the scope of the present article and will be dealt with in a future work.

5.3.2 A consistency result

The goal of this subsection is to display results of statistical nature, so that the TreeRank Tournament algorithm can be grounded in a strong validity framework.

Theorem 5.3.1. (Consistency) *For each $n \geq 1$, we consider scoring functions s_n, associated to the partition $\mathcal{F}_n$ of $\mathcal{X}$, resulting from a run of* TreeRank Tournament *algorithm in the case where $\mathcal{C}$ is a union stable of subset of $\mathcal{C}_n$. We assume that :*

- *$\mathcal{X}$ is bounded,*
- *$\mathcal{C}_n$ is union stable,*

- *the classes* $\mathcal{C}_n$ *are such that*

$$\lim_{n\to\infty} \frac{\log(\mathcal{S}(\mathcal{C}_n, n))}{n} = 0,$$

  where $\mathcal{S}(\mathcal{C}_n, n)$ *denotes the n-th shattering coefficient of the class of sets* $\mathcal{C}_n$.

- *the diameter of any cell of* $\mathcal{F}_n$ *goes to 0 when n tends to infinity.*

Then we have, as $n \to \infty$

$$\mathrm{VUS}(\mathrm{s}^*) - \mathrm{VUS}(\mathrm{s_n}) \to 0 \text{ almost surely.}$$

If, in addition,

- *the density of the distributions* $F_1^* = F_{s^*,1}$ *and* $F_3^* = F_{s^*,3}$ *are bounded,*
- *there exists a constant* $c > 0$ *such* $F_1^{*'}(u) > 1/c$ *and* $F_3^{*'}(u) > 1/c$ *for all* $u \in [0, 1]$,

then we have, as n goes to ∞,

$$d_\infty(s^*, (s_n)) \to 0 \text{ almost surely.}$$

The boundedness of $\mathcal{X}$ is a simplification which can be removed at the cost of a longer proof (the argument can be found in [Devroye *et al.*, 1996]). The complexity assumption is quite classical in machine learning, in particular in the case of empirical risk minimizer algorithm. This assumption controls the complexity of the partitions resulting from the TreeRank Tournament algorithm. This relies on the control of the n-th shattering coefficient of the collection of sets that can be obtained by union of sets $C \in \mathcal{C}_n$. Using the union stability assumption, this coefficient reduces to $\mathcal{S}(\mathcal{C}_n, n)$.

5.3.3 Learning rate bounds

The following noise assumption, used in [Clémençon *et al.*, 2013b] and generalizing that introduced in [Clémençon *et al.*, 2008] in the bipartite setup, shall be involved in the analysis. We recall the Assumption 2.

For $k \in \{1, 2\}$, the (pairwise) posterior probability given by $\eta_{k+1}(X)/(\eta_k(X) + \eta_{k+1}(X))$ is a continuous random variable and there exist $c < \infty$ and $a \in (0, 1)$ such that

$$\forall x \in \mathcal{X}, \ \mathbb{E}\left[\left|\frac{\eta_{k+1}(X)}{\eta_{k+1}(X) + \eta_k(X)} - \frac{\eta_{k+1}(x)}{\eta_{k+1}(x) + \eta_k(x)}\right|^{-a}\right] \leq c \ .$$

As revealed by the theorem below, equipped with this additional hypothesis, one may connect the performance of the splitting rule winner of the empirical tournament to that of the winner of the tournament based on the true VUS increment. The result is then established by following line by line the argument of Theorem 15 in [Clémençon & Vayatis, 2009b], see the sketch of proof given in the Appendix section.

Theorem 5.3.2. *Assume that* **Assumptions 1-4** *hold. Suppose that the class* $\mathcal{L}(\mathcal{X})$ *of subsets candidates is of finite* VC *dimension* V*, contains all level sets* $\{x \in \mathcal{X} : \eta(x) \geq t\}$*,* $t \in \mathbb{R}$*, of the regression function (or of optimal scoring functions equivalently) and that* $\mathcal{L}(\mathcal{X}) \cap \mathcal{C} = \mathcal{L}(\mathcal{C})$ *for all* $\mathcal{C} \in \mathcal{L}(\mathcal{X})$*. Then, there exists a constant* c_0 *and universal constants* c_1 *and* c_2 *such that, for all* $\delta > 0$*, with probability at least* $1 - \delta$*, we have: for all* $J \geq 1$ *and* $n \geq 1$*,*

$$d_1(s_{2^J}, s^*_{2^J}) \leq c_0^J \left\{ \left(c_1^2 V/n\right)^{\frac{a^J}{2(1+a)^J}} + \left(c_2^2 \log \delta/n\right)^{\frac{a^J}{2(1+a)^J}} \right\}.$$

Combined with Proposition 5.2.2, the result stated above provides rate bounds in the ROC space. Naturally, because of the hierarchical structure of the oriented partition produced by the TreeRank Tournament algorithm, slow rate bounds were expected. We point out however that the bounds exhibited hold true under very general assumptions and correspond to confidence regions in sup norm (analogous results in terms of VUS immediately follow).

5.4 Conclusion

The multipartite ranking problem is characterized by its global nature which is well reflected by function-like optimization criteria such as the ROC surface. The present chapter investigates an algorithm which iteratively builds a piecewise scoring function with a tree-structured partition over the input space. The splitting task is solved by organizing a tournament between optimal splitting rule for the bipartite sub-problems. We show that the proposed approximation scheme has a classical rates of convergence. We also carry a theoretical analysis of the empirical version of this approximation scheme and we state a consistency result in L_∞-norm. To our knowledge, this is the first result of this nature in multipartite ranking. The numerical performances of this algorithm are discussed in the next chapter.

5.5 Proofs

Proof of Theorem 5.2.2

We now show that the recursive approximation procedure introduced in Section 5.2 provides a sequence of piecewise scoring functions $(s_D)_{D \geq 0}$ with $N = 2^D$ constant parts which achieves an approximation error rate for the VUS of the order 2^{-2D}.

For any $(\alpha, \gamma) \in (\alpha^*_{D,k}, \alpha^*_{D,k+1}) \times (\gamma^*_{D,l+1}, \gamma^*_{D,l})$ we have, for any optimal scoring function s^*, by concavity :

$$\begin{aligned} \mathrm{ROC}(\mathrm{s}^*, \alpha, \gamma) - \mathrm{ROC}(\mathrm{s_D}, \alpha, \gamma) \leq & -\frac{1}{8}\left((\alpha^*_{\mathrm{D,k+1}} - \alpha^*_{\mathrm{D,k}})^2 \frac{\partial^2}{\partial^2 \alpha} \mathrm{ROC}(\mathrm{s}^*, \alpha, \gamma) \right) \\ & - \frac{1}{8}\left((\gamma^*_{D,k+1} - \gamma^*_{D,k})^2 \frac{\partial^2}{\partial^2 \gamma} \mathrm{ROC}(\mathrm{s}^*, \alpha, \gamma) \right) \end{aligned}$$

By assumption, the second derivatives of the optimal ROC surface are bounded and hence it suffices to check that for some constant C, we have

$$\forall k, \alpha^*_{D,k+1} - \alpha^*_{D,k} \leq C2^{-D} \text{ and } \gamma^*_{D,k} - \alpha^*_{D,k+1} \leq C2^{-D}$$

These inequalities follow immediately from a recurrence based on the next lemmas.

Lemma 5.5.1. *Consider $f : [0,1] \to [0,1]$ a differentiable and decreasing scoring function such that $m \leq f' \leq M < 0$. Take $x_0 < x_* < x_1$ such that $\max\{|x_0 - x_*|, |x_1 - x_*|\} \leq C|x_1 - x_0|$ with $C < 1$. Then, we have*

$$\max\{|f(x_0) - f(x_*)|, |f(x_1) - f(x_*)|\} \leq C'|f(x_1) - f(x_0)|$$

with $C' = 1 - (1 - C)M/m$

Proof. Using the theorem of finite increment to f, we have

$$f(x_0) - f(x_*) = (x_0 - x_*)f'(c)$$

and

$$f(x_0) - f(x_1) = (x_0 - x_1)f'(c').$$

Now, we need to use that the derivative is bounded and that there exist m and m such as $m \leq f' \leq M < 0$. Notice also that we have $|x_0 - x_*| > (1 - C)|x_0 - x_1|$. Combining these two properties, we obtain

$$\begin{aligned} f(x_0) - f(x_*) = (x_0 - x_*)f'(c) &\geq (x_0 - x_*)M \\ &\geq (1 - C)(x_0 - x_1)M \geq (1 - C)M/m(f(x_0) - f(x_1)) \end{aligned}$$

With the same argument we obtain $f(x_*) - f(x_1) \geq (1 - C)M/m(f(x_0) - f(x_1))$ and we obtain the desired result. □

Lemma 5.5.2. *Consider $f : [0,1] \to [0,1]$ a twice differentiable, decreasing and concave scoring function such that $m \leq f' \leq M < 0$. Take $x_0 < x_1$ and set x_* such that $f'(x_*) = (f(x_1) - f(x_0))/(x_1 - x_0)$. Then, we have*

$$\max\{x_0 - x_*|, |x_1 - x_*|\} \leq C|x_1 - x_0|$$

with $C = 1 - M/2m$

Proof. As f' is continuous and decreasing, we can use the following expression $x_* = f'^{-1}((f(x_1) - f(x_0))/(x_1 - x_0))$. By applying the theorem of finite increment to f'^{-1} between $f'(x_*)$ and $f'(x_1)$, we have

$$x_* - x_1 = (f'(x_*) - f'(x_1))(f'^{-1})'(c)$$

for some c. So we deduce that

$$x_* - x_0 = x_1 - x_0 + (f'(x_*) - f'(x_1))(f'^{-1})'(c).$$

The Taylor formula says that

$$f(x_0) = f(x_1) + (x_0 - x_1)f'(x_1) + (x_0 - x_1)^2 f''(c')/2$$

and we deduce that $f'(x_*) - f'(x_1) = (x_0 - x_1)f''(c')/2$. Using that $(f'^{-1})'(c) = 1/f''(f'^{-1}(c))$ and $m \leq f' \leq M < 0$, we obtain

$$(x_* - x_0) \leq x_1 - x_0 + (x_0 - x_1)M/2m.$$

Using the same arguments to $x_1 - x_*$ leads to the desired result. □

Proof of Theorem 5.3.1

(L_1 metric). Using Theorem 2 in [Clémençon *et al.*, 2013b], we have

$$\mathrm{VUS}(\mathrm{s}^*) - \mathrm{VUS}(\hat{\mathrm{s}}) \leq \left(\mathrm{AUC}^*_{\phi_1,\phi_2} - \mathrm{AUC}_{\phi_1,\phi_2}(\hat{\mathrm{s}})\right) + \left(\mathrm{AUC}^*_{\phi_2,\phi_3} - \mathrm{AUC}_{\phi_2,\phi_3}(\hat{\mathrm{s}})\right).$$

We recall that

$$\mathrm{AUC}_{\phi_1,\phi_2}(\hat{\mathrm{s}}) = \mathbb{P}\{\hat{\mathrm{s}}(\mathrm{X}) < \hat{\mathrm{s}}(\mathrm{X}')|\mathrm{Y} = 1, \mathrm{Y}' = 2\} + \frac{1}{2}\sum_{\mathrm{j}=0}^{2^{\mathrm{D}}-1} \alpha(\mathrm{C}_{\mathrm{d,j}})\beta(\mathrm{C}_{\mathrm{d,j}}).$$

Now, mimicking the proof of Proposition 13 in [Clémençon *et al.*, 2013b], we have

$$\mathrm{AUC}^*_{\phi_1,\phi_2} - \mathrm{AUC}_{\phi_1,\phi_2}(\hat{\mathrm{s}}) \leq \frac{\mathbb{E}[|(\eta_1(\mathrm{X}) - \hat{\eta}_1)(\mathrm{X})| + |(\eta_2(\mathrm{X}) - \hat{\eta}_2(\mathrm{X}))|]}{\mathrm{p_1 p_2}} + \max_{0\leq \mathrm{j} \leq 2^{\mathrm{D}}-1} \beta(\mathrm{C}_{\mathrm{D,j}})$$

The term on the right hand side of the equation vanishes as $n \to \infty$ by assumption, while the first term can be handled by reproducing exactly the argument of Theorem 21.2 from [Devroye *et al.*, 1996].

(L_∞ supnorm metric). First we introduce the notation :

$$\hat{F}^*_1(t) = \frac{1}{n_1}\sum_{i=1}^{n} \mathbb{I}\{\hat{\eta}_1(X_i) \leq t, Y_i = 1\}$$

and observe that under our assumptions,

$$\hat{F}^*_1(\hat{\eta}(x)) = \sum_{l=0}^{2^D-1} \hat{\alpha}(C_{D,l})\mathbb{I}\{x \in R_{D,l}\}.$$

Notice that :

$$F_{\hat{s},1}(\hat{s}(x)) = \sum_{l=0}^{2^D-1} \alpha(C_{D,l})\mathbb{I}\{x \in R_{D,l}\}.$$

Using Proposition 5 in [Clémençon *et al.*, 2013b], we have for any $(\alpha, \gamma) \in [0,1]^2$:

$$\mathrm{ROC}^*(\mathrm{s}^*, \alpha, \gamma) - \mathrm{ROC}(\hat{\mathrm{s}}, \alpha, \gamma) \leq (\Theta_1(\hat{\mathrm{s}}, \alpha) + \Theta_2(\hat{\mathrm{s}}, \gamma)),$$

where

$$\begin{aligned}
\Theta_1(s,\alpha) &= \frac{\mathbb{I}\{\alpha \neq 0\}}{p_2 Q^{(1)}(\eta_1, \alpha)} \mathbb{E}\left[\left|\eta_1(X) - Q^{(1)}(\eta_1, \alpha)\right| \cdot \mathbb{I}\{R^{(1)}_{s^*,\alpha} \Delta R^{(1)}_{\hat{s},\alpha}\}\right] \\
&\quad + \frac{1}{p_2 Q^{(1)}(\eta_1, \alpha)} (\alpha - 1 + F_{\hat{s},1}(Q(\hat{s}, \alpha))), \\
\Theta_2(\hat{s},\gamma) &= \frac{\mathbb{I}\{\gamma \neq 1\}}{p_2 Q^{(3)}(\eta_3, 1-\gamma)} \mathbb{E}\left[\left|\eta_3(X) - Q^{(3)}(\eta_3, 1-\gamma)\right| \cdot \mathbb{I}\{R^{(3)}_{s^*,1-\gamma} \Delta R^{(3)}_{s,1-\gamma}\}\right] \\
&\quad + \frac{1}{p_2 Q^{(3)}(\eta_3, \alpha)} (1 - \gamma - F_{\hat{s},3}(Q(\hat{s}, 1-\gamma))) \ .
\end{aligned}$$

Recall that we use the notation $A \Delta B = (A \setminus B) \cup (B \setminus A)$ for the symmetric difference between sets A and B. For the second term of $\Theta_1(s,\alpha)$, noticing that if $\alpha(R_{D,l}) \leq \alpha \leq \alpha(R_{D,l+1})$, then

$$0 \leq \alpha - 1F_{\hat{s},1}(Q(\hat{s}, \alpha)) \leq \alpha(C_{D,l}).$$

We then use the assumption that the cells of the partitions $\mathcal{F}_n$ tend to zero when n grows to infinity. Now, for the first term of $\Theta_1(s,\alpha)$, we have

$$\mathbb{E}\left[\left|\eta_1(X) - Q^{(1)}(\eta_1, \alpha)\right| \cdot \mathbb{I}\{R^{(1)}_{s^*,\alpha} \Delta R^{(1)}_{\hat{s},\alpha}\}\right] \leq c\mathbb{E}\left[|F_1^*(\eta_1(X) - (1-\alpha)| \cdot \mathbb{I}\{R^{(1)}_{s^*,\alpha} \Delta R^{(1)}_{\hat{s},\alpha}\}\right]$$

by virtue of the theorem of finite increment. We easily see that for $x \in R^{(1)}_{s^*,\alpha} \Delta R^{(1)}_{\hat{s},\alpha}$

$$|F_1^*(\eta_1(x) - (1-\alpha)| \leq |F_1^*(\eta_1(x)) - F_{\hat{s},1}(\hat{s}(x))| \,.$$

Now this last term can be decomposed as follows :

$$\begin{aligned}
|F_1^*(\eta_1(x)) - F_{\hat{s},1}(\hat{s}(x))| \leq |F_1^*(\eta_1(x)) - F_1^*(\hat{\eta}_1(x))| + |F_1^*(\hat{\eta}_1(x)) - \hat{F}_1^*(\hat{\eta}_1(x))| \\
+ |\hat{F}_1^*(\hat{\eta}_1(x)) - F_{\hat{s},1}(\hat{s}(x))|.
\end{aligned}$$

The first term can be handled by combining the finite increments theorem and the theorem 21.2 in [Devroye *et al.*, 1996]. The middle term goes to zero by Glivenko-Cantelli theorem. The last term is controlled as soon as, almost surely, $\hat{\alpha}(C_{D,l}) - \alpha(C_{D,l})$ converges to zero, as n tends to infinity. Using again Glivenko-Cantelli leads to the desired result.

Proof of Theorem 5.3.2

We are going to show that $\forall j \in \{1, \ldots, J\}, l \in \{0, \ldots, 2^{J-1} - 1\}$, with probability $1 - \delta$,

$$\mathbb{E}\left[\mathbb{I}\{\mathcal{C}^*_{j,2l} \Delta \mathcal{C}_{j,2l}\}\right] \leq C \times B\left(\frac{(1+a)^{(j)}}{a^{(j)}}, n, \delta\right), \text{ and} \tag{5.3}$$

$$|\mathrm{VUS}(\mathrm{s}^*_{2^\mathrm{j}}) - \mathrm{VUS}(\hat{\mathrm{s}}_{2^\mathrm{j}})| \leq \mathrm{C} \times \mathrm{B}\left(\frac{(1+\mathrm{a})^\mathrm{j}}{\mathrm{a}^\mathrm{j}}, \mathrm{n}, \delta\right), \tag{5.4}$$

where $B(d,n,\delta) = \left(\frac{c_1^2 V}{n}\right)^{\frac{1}{2d}} + \left(\frac{c_2^2 \log\delta}{n}\right)^{\frac{1}{2d}}$

First, we detail what happen for the first step. By symmetry, we can suppose that $\mathcal{C}_{1,0} = \widetilde{\mathcal{C}}^{(1)}$. Using Lemma 19 in [Clémençon & Vayatis, 2009b], we have, with probability $1-\delta$, $\mathrm{AUC}_{1,2}(\mathrm{s}^*_1) - \mathrm{AUC}_{1,2}(\hat{\mathrm{s}}_1) \leq \kappa_1 \mathrm{B}(1,\mathrm{n},\delta)$. We can easily show that we have $\mathrm{AUC}_{2,3}(\mathrm{s}^*_2) - \mathrm{AUC}_{2,3}(\hat{\mathrm{s}}_2) \leq \frac{\mathrm{p_3 p_2}}{2(\mathrm{p_3}+\mathrm{p_2})}\mathbb{E}\left[\mathbb{I}\{\mathrm{X} \in \mathcal{C}_{1,0}\Delta\mathcal{C}^*_{1,0}\}\right]$. Combining the noise condition 2 and the Hölder inequality, we obtain that $\mathbb{E}\left[\mathbb{I}\{X \in \mathcal{C}_{1,0}\Delta\mathcal{C}^*_{1,0}\}\right]$ is upper bounded by

$$\left(\frac{2(p_1+p_2)}{p_1 p_2}\mathrm{AUC}_{1,2}(\mathrm{s}^*_2) - \mathrm{AUC}_{1,2}(\hat{\mathrm{s}}_2)\right)^{\frac{a}{1+a}} \times c^{\frac{1}{1+a}}.$$

Finally, using that $|\mathrm{VUS}(\mathrm{s}^*_2) - \mathrm{VUS}(\hat{\mathrm{s}}_2)| \leq |\mathrm{AUC}_{1,2}(\mathrm{s}^*_2) - \mathrm{AUC}_{1,2}(\hat{\mathrm{s}}_2)| + |\mathrm{AUC}_{2,3}(\mathrm{s}^*_2) - \mathrm{AUC}_{2,3}(\hat{\mathrm{s}}_2)|$, we have, with probability $1-\delta$, $|\mathrm{VUS}(\mathrm{s}^*_2) - \mathrm{VUS}(\hat{\mathrm{s}}_2)|$ upper bounded by $C.B((1+a)/a, n, \delta)$.

Let $j > 1$. Suppose that the bound stated in 5.3 holds for all $j-1$.
We have $\mathrm{VUS}(\mathrm{s}^*_{2^\mathrm{j}}) - \mathrm{VUS}(\mathrm{s}_{2^\mathrm{j}}) \leq |\mathrm{AUC}_{1,2}(\mathrm{s}^*_{2^\mathrm{j}}) - \mathrm{AUC}_{1,2}(\hat{\mathrm{s}}_{2^\mathrm{j}})| + |\mathrm{AUC}_{2,3}(\mathrm{s}^*_{2^\mathrm{j}}) - \mathrm{AUC}_{2,3}(\hat{\mathrm{s}}_{2^\mathrm{j}})|$. Using the bound in [Clémençon & Vayatis, 2009b], we have $2|\mathrm{AUC}_{1,2}(\mathrm{s}^*_{2^\mathrm{j}}) - \mathrm{AUC}_{1,2}(\hat{\mathrm{s}}_{2^\mathrm{j}})| \leq \sum_{\mathrm{l}=1}^{2^{\mathrm{j}-1}-1} |\mathrm{F}_1(\mathcal{C}^*_{\mathrm{j}-1,\mathrm{l}})\mathrm{F}_2(\mathcal{C}^*_{\mathrm{j}-1,\mathrm{l}})\Lambda_{1,2}(\mathcal{C}^*_{\mathrm{j},2\mathrm{l}} \mid \mathcal{C}^*_{\mathrm{j}-1,\mathrm{l}}) - \mathrm{F}_1(\mathcal{C}_{\mathrm{j}-1,\mathrm{l}})\mathrm{F}_2(\mathcal{C}_{\mathrm{j}-1,\mathrm{l}})\Lambda_{1,2}(\mathcal{C}_{\mathrm{j},2\mathrm{l}} \mid \mathcal{C}_{\mathrm{j}-1,\mathrm{l}})|$. By symmetry, we suppose that the tournament have chosen $\mathcal{C}_{j,2l} = \arg\max_{\mathcal{C}\in\mathcal{C}_{j-1,l}} \tilde{\Lambda}_{1,2}(\mathcal{C}|\mathcal{C}_{j-1,l})$, i.e. the solution of the problem 1 vs 2. We introduce the set $\bar{\mathcal{C}}_{j,2l} = \arg\max_{\mathcal{C}\subset\mathcal{C}_{j-1,l}} \Lambda_{1,2}(\mathcal{C} \mid \mathcal{C}_{j-1,l})|$. We have

$$\begin{aligned}
&|F_1(\mathcal{C}^*_{j-1,l})F_2(\mathcal{C}^*_{j-1,l})\Lambda_{1,2}(\mathcal{C}^*_{j,2l} \mid \mathcal{C}^*_{j-1,l}) - F_1(\mathcal{C}_{j-1,l})F_2(\mathcal{C}_{j-1,l})\Lambda_{1,2}(\mathcal{C}_{j,2l} \mid \mathcal{C}_{j-1,l})| \\
&\leq |F_1(\mathcal{C}^*_{j-1,l})F_2(\mathcal{C}^*_{j,2l}) - F_2(\mathcal{C}^*_{j-1,l})F_1(\mathcal{C}^*_{j,2l}) - F_1(\mathcal{C}_{j-1,l})F_2(\bar{\mathcal{C}}_{j,2l}) + F_2(\mathcal{C}_{j-1,l})F_1(\bar{\mathcal{C}}_{j,2l})| \\
&+|F_1(\mathcal{C}_{j-1,l})F_2(\bar{\mathcal{C}}_{j,2l}) + F_2(\mathcal{C}_{j-1,l})F_1(\bar{\mathcal{C}}_{j,2l}) - F_1(\mathcal{C}_{j-1,l})F_2(\mathcal{C}_{j,2l}) + F_2(\mathcal{C}_{j-1,l})F_1(\mathcal{C}_{j,2l})| = A_{j,2l} + B_{j,2l}.
\end{aligned}$$

Using the VC inequality as for the first split, with probability $1-\delta$, the quantity $B_{j,2l}$ is bounded by $B((1+a)/a, n, \delta)$. Notice in particular that we have, with probability $1-\delta$, $\mathbb{E}\left[\mathbb{I}\{X \in \bar{\mathcal{C}}_{j,2l}\Delta\mathcal{C}_{j,2l}\}\right] \leq C.B((1+a)/a, n, \delta)$.

Reproducing the argument of [Clémençon & Vayatis, 2009b], $A_{j,2l}$ is bounded by $|F_1(\mathcal{C}^*_{j-1,l}) - F_1(\mathcal{C}_{j-1,l})| + |F_2(\mathcal{C}^*_{j-1,l}) - F_2(\mathcal{C}_{j-1,l})| \leq B\left(\frac{(1+a)^{(j-1)}}{a^{(j-1)}}, n, \delta\right)$ using the inequality 5.3. Now, we have to bound $\mathbb{E}\left[\mathbb{I}\{\mathcal{C}^*_{j,2l}\Delta\mathcal{C}_{j,2l}\}\right]$. We have $\mathbb{E}\left[\mathbb{I}\{\mathcal{C}^*_{j,2l}\Delta\mathcal{C}_{j,2l}\}\right] \leq \mathbb{E}\left[\mathbb{I}\{\mathcal{C}^*_{j,2l}\Delta\bar{\mathcal{C}}_{j,2l}\}\right] + \mathbb{E}\left[\mathbb{I}\{\bar{\mathcal{C}}_{j,2l}\Delta\mathcal{C}_{j,2l}\}\right]$. Using the Hölder inequality and the noise condition 2, we have

$$\mathbb{E}\left[\mathbb{I}\{\mathcal{C}^*_{j,2l}\Delta\bar{\mathcal{C}}_{j,2l}\}\right] \leq C|A_{j,2l}|^{\frac{a}{1+a}} \leq C.B\left(\frac{(1+a)^j}{a^j}, n, \delta\right).$$

Thus we obtain the inequality 5.3. From that inequality, we can easily deduce the inequality 5.4.

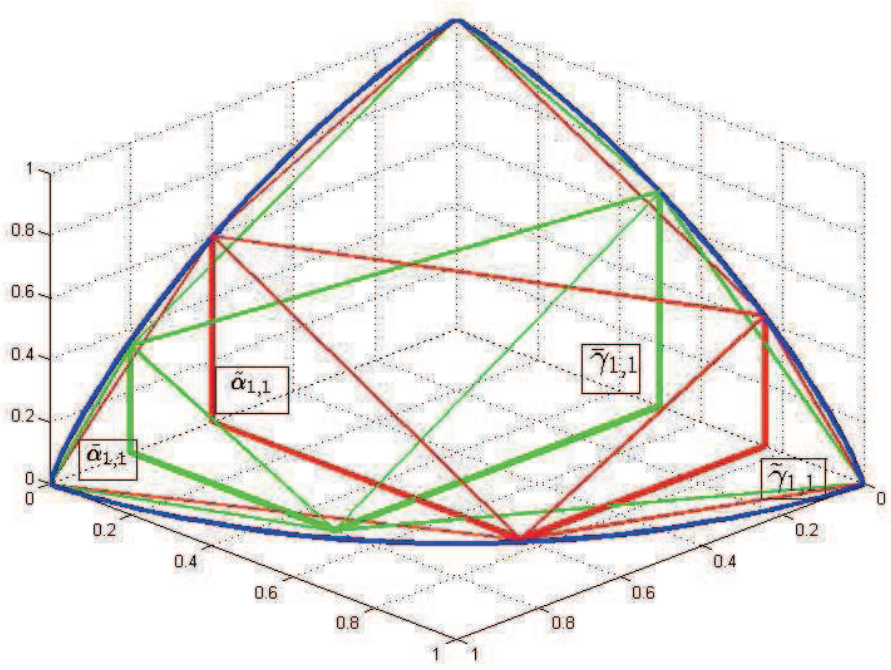


Figure 5.1: First split : *tournament* to maximize the VUS increase

TreeRank Tournament

1. (Input.) Training sample $\mathcal{D}_n$, LeafRank algorithm $\mathcal{L}$, ranking tree depth J.

2. (Initialization.) Set $\mathcal{C}_{0,0} = \mathcal{X}$ and $s_0(x) \equiv 1$.

3. (Iterations.) For $m = 1, \ldots, 2^J$, define $j = \langle \log m / \log 2 \rangle$ and $l = m - 2^j$, and then

   (a) (LeafRank runs.) For all $k \in \{1, 2\}$, run algorithm $\mathcal{L}$ in order to output

   $$\widetilde{\mathcal{C}}^{(k)} = \underset{\mathcal{C} \in \mathcal{L}(\mathcal{C}_{j,l})}{\arg\max} \widehat{\Lambda}_{k,k+1}(\mathcal{C} \mid \mathcal{C}_{j,l}).$$

   (b) (Tournament.) Compute

   $$\mathcal{C}_{j+1,2l} = \underset{\widetilde{\mathcal{C}}^{(k)},\ k=1,\ 2}{\arg\max} \widehat{\mathrm{VUS}}_{\mathcal{C}_{d,l}}(\widetilde{\mathcal{C}}^{(k)}),$$

   and set $\mathcal{C}_{j+1,2l+1} = \mathcal{C}_{j,l} \setminus \mathcal{C}_{j+1,2l}$.

4. (Output.) Compute the piecewise constant scoring function :

   $$s_{2^J}(x) = \sum_{l=0}^{2^J-1} (2^J - l) \cdot \mathbb{I}\{x \in \mathcal{C}_{J,l}\}.$$

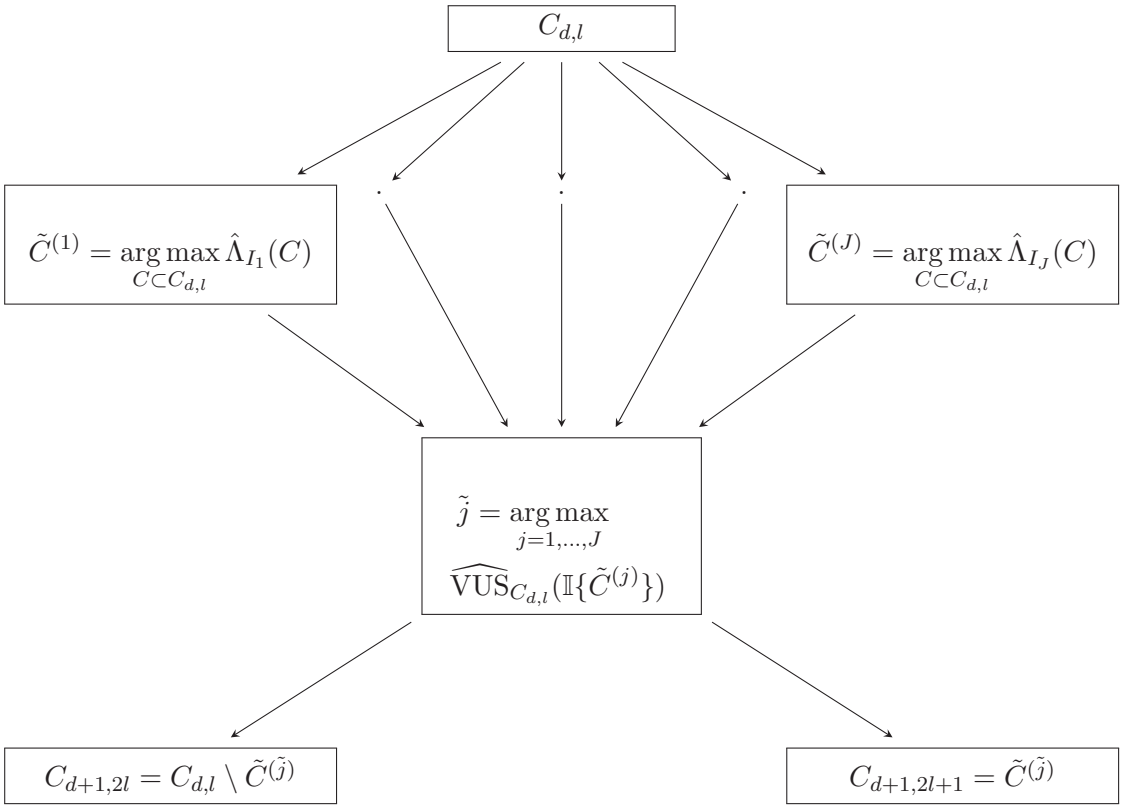


Figure 5.2: Splitting the $C_{d,l}$ set.

Chapter 6

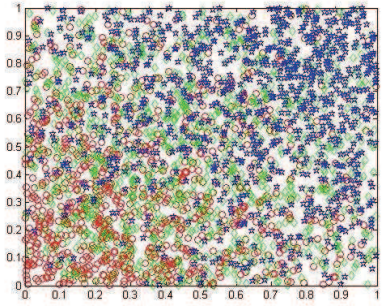
Numerical experiments

The purpose of this chapter is to illustrate the approaches described in the previous chapters by numerical results and provide some empirical evidence for their efficiency. Our goal is to show that, beyond their theoretical validity, our algorithms work in practice and to provide a detailed empirical study of their performance on benchmark artificial/real datasets compared to that of possible competitors.
First the methodology to measure the performance of scoring functions is described. Several datasets with different features are considered, in particular the number of labels and the size of the dataset. Moreover, we simulate a dataset with supports issue in order to see the influence over the decompositions of the multipartite ranking into bipartite ranking sub-problems. The results are viewed through the estimation of the VUS. Since in applications such as information retrieval the top of the ranking is more important, we compute a performance measure of the quality of the top of the ranking.
This methodology is used to give insights over the procedure of Chapters 4 and 5. Several sets-up of the procedureare compared. Then, we choose the procedures that give the best results overall the datasets in order to compare with the competitors. Description of the competitors that are RankBoost, SVMrank and regression least square (RLS) are provided. All these procedures are based on the same principle : use a classification procedure on pairs of observation in order to learn scoring functions. The performances of the scoring function depend a lot of the dataset that permit us to understand when the TreeRank procedures are a valid choice.

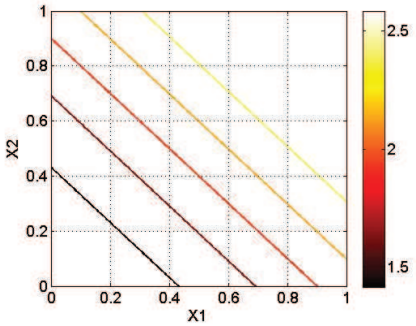
The rest of the chapter is structured as follows. In section 6.1, we present all the datasets we use for the numerical experiments. For each simulated distribution, we provide a sample as well as an optimal scoring function. In section 6.2, we present the criteria that are used to evaluate the scoring functions. In section 6.3, we present how we implement the methods introduced in this manuscript and we compare the numerical performances. In section 6.4, we present in detail the competitors and the implementation we use. The results are given through a visualization bar and the exact values can be found in the Annex 6.5.

6.1 Data description

In this section, we describe the dataset that are used to compare the performance and the algorithms.



a. Pooled sample: red circles represent instances with label "1", green diamonds those with label "2" and blue stars those with label "3"



b. Optimal level sets.

Figure 6.1: First example - Mixture of Gaussian distributions

6.1.1 Simulated data

Mixtures of Gaussian distributions.

Consider Z a q-dimensional random vector from a Gaussian distribution drawn $\mathcal{N}(\mu, \Gamma)$, and a Borelian set $C \subset \mathbb{R}^q$. We denote by $\mathcal{N}_C(\mu, \Gamma)$ the conditional distribution of Z given $Z \in C$. Equipped with this notation, we can write the class distributions used in this example as:

$$\begin{aligned}
\phi_1(x) &= \mathcal{N}_{[0,1]^2}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1/4 & 0 \\ 0 & 1/4 \end{pmatrix}\right) \\
\phi_2(x) &= \mathcal{N}_{[0,1]^2}\left(\begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}, \begin{pmatrix} 1/4 & 0 \\ 0 & 1/4 \end{pmatrix}\right) \\
\phi_3(x) &= \mathcal{N}_{[0,1]^2}\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1/4 & 0 \\ 0 & 1/4 \end{pmatrix}\right)
\end{aligned}$$

When $p_1 = p_2 = p_3 = 1/3$, the regression function is then an increasing transform of $(x_1, x_2) \in [0,1]^2 \mapsto x_1 + x_2$, given by:

$$\eta(x) = \frac{2.79 \cdot e^{-(x_1+x_2)^2} + 2 \cdot 1.37 \cdot e^{-(x_1+x_2-1)^2} + 3 \cdot 2.79 \cdot e^{-(x_1+x_2-2)^2}}{2.79 \cdot e^{-(x_1+x_2)^2} + 1.37 \cdot \exp^{-(x_1+x_2-1)^2} + 2.79 \cdot e^{-(x_1+x_2-2)^2}}.$$

The simulated dataset is plotted in Fig. 6.1a, while some level sets of the regression function are represented in 6.1b. We have drawn a sample of size $n = 3000$.

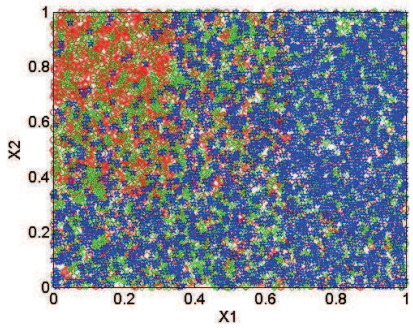
Mixtures of uniform distributions.

The artificial data sample used in this second example is represented in Fig 6.2.a and has been generated as follows. The unit square $\mathcal{X} = [0,1]^2$ is split into 9 squares of equal size and the scoring function s^* is defined as the constant function on each of these squares depicted by Fig. 6.2.b. We then choose the uniform distribution over

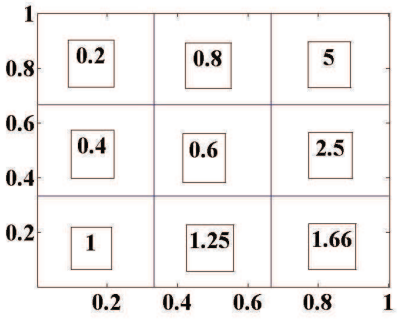
Table 6.1: Values of the η_k's on each of the nine subsquares of $[0,1]^2$, *cf* Fig. 6.2 b

s^*	$s^*_{1,2}$	$s^*_{2,3}$	η_1	η_2	η_3
0.2	0.2	0.2	0.7692	0.2000	0.0308
0.4	0.4	0.2	0.6250	0.3250	0.0500
0.6	0.8	0.6	0.3968	0.4127	0.1905
0.8	0.8	0.8	0.3731	0.3881	0.2388
1	1	1	0.3030	0.3939	0.3030
1.25	1.25	1	0.2581	0.4194	0.3226
1.66	1.66	1.66	0.1682	0.3645	0.4673
2.5	2.5	2.5	0.0952	0.3095	0.5952
5	2.5	5	0.0597	0.1940	0.7463

the unit square as marginal distribution and take $\Phi_{2,1}(x) = s^*_{1,2}(x)/1.3$ and $\Phi_{3,2} = 1.3 \times s^*_{2,3}(x)$. As $s^*_{1,2}$ and $s^*_{2,3}$ are non-decreasing functions of s^* (see Table 6.1), $\Phi_{2,1}$ and $\Phi_{3,2}$ are thus non-decreasing functions of s^*, by virtue of Theorem 1.1.1, and the class distributions checks the monotonicity assumption 1. Computation of the η_i's on each part of $\mathcal{X}$ is then straightforward, see Table 6.1.



a. Pooled sample: red circles represent instances with label "1", green diamonds those with label "2" and blue stars those with label "3"



b. Optimal scoring function s^*.

Figure 6.2: Second example - Mixtures of uniform distributions.

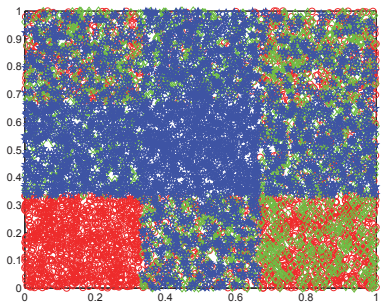
Mixture of uniform distributions with supports issue.

For this experiment, we use a mixture of uniform distributions on the unit square $[0,1]^2$. Specifically, it is divided into 9 equal parts and we put the best score s^* as shown in Figure 6.3 b. In table 6.2 are given the values of conditional distributions and the values of functions $s^*_{1,2}$ and $s^*_{2,3}$ that are increasing transformations of s^*, so the assumption 1 is verified. In Figure 6.3 a, we present a sample of 9000 observations

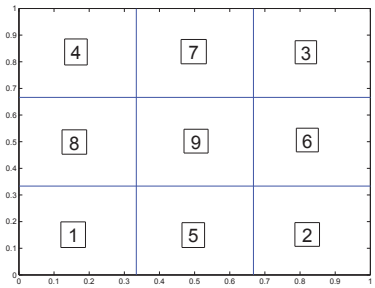
Table 6.2: Values of the η_k's on each of the nine subsquare of $[0,1]^2$, *cf* Fig. 6.3 b

s^*	$s^*_{1,2}$	$s^*_{2,3}$	η_1	η_2	η_3
0.2	0	0	1	0	0
0.4	0.4	0	0.7143	0.2857	0
0.6	0.8	0.6	0.4386	0.3509	0.2105
0.8	1	0.8	0.3571	0.3571	0.2857
1	1.25	1	0.2857	0.3571	0.3571
1.25	2.5	1	0.1667	0.4167	0.4167
1.66	5	1.66	0.0698	0.3488	0.5814
2.5	$+\infty$	2.5	0	0.2857	0.7143
5	$+\infty$	$+\infty$	0	0	1

following the distributions mixture.



a. Pooled sample: red circles represent instances with label "1", green diamonds those with label "2" and blue stars those with label "3"



b. Optimal scoring function s^*.

Figure 6.3: Mixtures of uniform distributions with supports issue

6.1.2 Real dataset

We also illustrate the methodology promoted in this thesis by implementing it on a real data set, the *Cardiotocography Data Set* considered in [Frank & Asuncion, 2010] namely. The data have been collected as follow: 2126 fetal cardiotocograms (CTG's in abbreviated form) have been automatically processed and the respective diagnostic features measured. The CTG's have been next analyzed by three expert obstetricians and a consensus ordinal label has been then assigned to each of them, depending on the degree of anomaly observed: 1 for "normal", 2 for "suspect" and 3 for "pathologic".

We also carried out experiments based on four datasets with ordinal labels (ERA,

Table 6.3: Description of the simulated datasets

Name	sample size	feature space dimension	number of classes	supports issue
Gauss2d	3000	2	3	No
Unif2d	9000	2	3	No
Unif2dsi	9000	2	3	Yes

Table 6.4: Description of the real datasets

Name	sample size	features space dimension	number of classes
Cardio	2126	20	3
ERA 1-9	1000	4	9
ERA 1-7	951	4	7
ESL 3-7	451	4	9
LEV 0-4	1000	4	5
LEV 0-3	973	4	4
SWD 2-5	1000	10	4
SWD 3-5	978	10	3
MQ2008	15211	46	3

ESL, LEV and SWD namely), considered in [David, 2008]. Because of the wide disparity between some class sizes, data with certain labels are ignored (in the ESL dataset for instance, the class "1" counts only two observations).

We also implemented the approach promoted in this paper on the benchmark LETOR datasets, (see `research.microsoft.com/en-us/um/people/letor/`), by means of the same ranking algorithm as that used in the previous experiment. To be more precise, we used the query set MQ2008, where pairs "page-query" assigned to a discrete label ranging from 0 to 2 (*i.e.* "non-relevant" - "relevant" - "extremely relevant") are gathered. In the dataset, 46 features are collected over 15 211 instances in MQ2008. All the characteristics of the datasets are summed up in the Table 6.4.

6.2 Criteria

In all the experiments, we evaluate the accuracy of the methods using the VUS criterion. For each experiment, we estimate the VUS by a cross validation procedure. Specifically, for a data set $\mathcal{D}_n$ we create a collection of $V = 5$ sub-samples denoted

$\{\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(V)}\}$ in order to estimate the empirical $\overline{\mathrm{VUS}}^{(t)}$ given by the formula :

$$\overline{\mathrm{VUS}}^{(t)} = \frac{1}{V} \sum_{v=1}^{V} \widehat{\mathrm{VUS}}^{(v)},$$

where for all $v \in \{1, \ldots, V\}$, $\widehat{\mathrm{VUS}}^{v}$ is the empirical VUS associated to the scoring function $s^{(-v)}$, learned over the training dataset $\mathcal{D}^{(-v)} = \mathcal{D}_n \setminus \mathcal{D}^{(v)}$ and evaluate on the dataset $\mathcal{D}^{(v)}$. Moreover, we resample $B = 5$ times this procedure to produce more accurate estimators and we denote $\hat{\sigma}$ the standard deviation of the empirical VUS evaluated over the $B \times V$ iterations of the procedure.

We mention that alternative ranking performance statistics have been used in the multipartite ranking framework, namely the C-INDEX (see [Fürnkranz *et al.*, 2009], [Herbrich *et al.*, 2000]) and the JONCKHEERE-TERPSTRA STATISTIC (*JPstat* in abbreviated form, see [Hand & Till, 2001] and [Higgins, 2004]). The C-INDEX evaluate the probability that a scoring function ranks correctly two random pairs (X, Y) and (X', Y') i.e. $\mathbb{P}\{s(X) < s(X')|Y < Y'\}$ which is equal to

$$\frac{1}{\sum_{1 \leq k < k' \leq K} p_k p_{k'}} \sum_{1 \leq k < k' \leq K} p_k p_{k'} \mathrm{AUC}_{\phi_k, \phi_{k'}}(\mathrm{s}).$$

The *JPstat* is the empirical version of the mean of the AUC i.e.

$$\frac{2}{K(K-1)} \sum_{1 \leq k < k' \leq K} \mathrm{AUC}_{\phi_k, \phi_{k'}}(\mathrm{s}).$$

However, we find these criteria less discriminative than the VUS criterion so we do not report them in the result.

We also want to compare the top of the ranking but the VUS is a global criterion not adapted to this end. So we introduce a notion of local VUS mimicking the local AUC introduced in [Clémençon & Vayatis, 2007]. Recall that we denote $Q(s, u_0)$ the $(1 - u_0)$-quantile of the random variable $s(X)$. The probabilistic definition of the local VUS is given by the following equation :

$$\mathrm{LocVUS}(\mathrm{s}, \mathrm{u}_0) = \mathbb{P}\{\mathrm{s(X)} < \mathrm{s(X')} < \mathrm{s(X'')}, \mathrm{s(X'')} > \mathrm{Q(s, u_0)}|\mathrm{Y} = 1, \mathrm{Y'} = 2, \mathrm{Y''} = 3\}$$

where (X, Y), (X', Y'), (X'', Y'') are i.i.d. copies of distribution P. The empirical version of this criterion counts the number of triplets that are correctly ranked by the scoring function s when the highest value belongs to in the top-u_0% of the ranking. As for the empirical VUS, it is possible to implement a fast computation of the empirical LocVUS by modifying a bit the algorithm given in Annex 1.5.

6.3 TREERANK methods

First, we describe the algorithms based on the Kendall aggregation of scoring function and the TREERANK TOURNAMENT. A discussion over the results of the methods is given, highlighting the benefits and drawbacks.

6.3.1 Our algorithms in action

The learning algorithm used for solving the bipartite ranking subproblems at the first stage of the procedure is the TreeRank procedure based on locally weighted versions of the CART method (with axis parallel splits), see [Clémençon *et al.*, 2011a] for a detailed description of the algorithm (as well as [Clémençon & Vayatis, 2009b] for rigorous statistical foundations of this method). Precisely, we use a package from R statistical software (see http://www.r-project.org) implementing TreeRank (with the "default" parameters: minsplit = 1, maxdepth = 10, mincrit = 0), available at http://treerank.sourceforge.net, see [Baskiotis *et al.*, 2010]. We choose to use the SVM version of the LeafRank at each split of the TreeRank procedure because it can handle large datasets. We use linear kernel with parameter $C = 20$. To be more stable, we use the TreeRankForest implementation with 50 trees. The scoring rules produced at stage 1 are thus (tree-structured and) piecewise constant, making the aggregating procedure described in sub-subsection 4.1.3 quite feasible. Indeed, if $s_1, \ldots, s_M$ are scoring functions that are all constant on the cells of a finite partition $\mathcal{P}$ of the input space $\mathcal{X}$, one easily sees that the infimum $\inf_{s \in \mathcal{S}_0} \sum_{m=1}^{M} d_{\tau_\mu}(s, s_m)$ reduces to a minimum over a finite collection of scoring functions that are also constant on $\mathcal{P}$'s cells and is thus attained. As underlined in subsection 4.1.3, when the number of cells is large, median computation may become practically unfeasible and the use of a meta-heuristic can be then considered for approximation purpose (simulated annealing, tabu search, *etc.*), here the ranking obtained by taking the mean ranks over the $K-1$ rankings of the test data has been improved in the Kendall consensus sense by means of a standard simulated annealing technique. We recall that we note $\mathcal{D}_k = \{(x, y) \in \mathcal{D} | y = k\}$ the restriction of the dataset to the observations with label "k". The scoring function learned using the datasets $\mathcal{D}_k \cup \mathcal{D}_l$ with $k < l$ is called "TRkl" and these scoring functions are aggregated through the procedure described in sub-subsection 4.1.3, yielding the score called "AggRR" (as reference to "round robin"). Similarly, the scoring function learned using the datasets $\cup_{k'=1}^{k} \mathcal{D}_{k'}$ versus $\cup_{k'=k+1}^{K} \mathcal{D}_{k'}$ with $k \in \{1, \ldots, K\}$ is called "TRkFH" and these scoring functions are aggregated through the procedure described in sub-subsection 4.1.3, yielding the score called "AggFH" (as reference to [Frank & Hall, 2001]). Finally, the scoring function learned using the datasets $\cup_{k'=1}^{k} \mathcal{D}_{k'}$ versus $\cup_{k'=l}^{K} \mathcal{D}_{k'}$ with $k < l$ is called "TRklFH" and these scoring functions are aggregated through the procedure described in sub-subsection 4.1.3, yielding the score called "AggPFH" (as "pairwise" FH decomposition).

We also implement the modified version of the TreeRank algorithm presented in Chapter 5, called TreeRank Tournament. We recall that the difference with the original algorithm is that at each split of the TreeRank algorithm, we organize a competition between binary classifiers (i.e. classifier that discriminate two populations). Then we choose the classifier that maximizes the VUS on the data available at this step. As for the aggregation procedure, we have to decide which classifiers are in competition so we use the same decomposition of the problem. The

scoring function using the "round robin" decomposition is called "DTRRR", the one using the FH decomposition is called "DTRFH" and the one with the pairwise FH decomposition is called "DTRPFH". To implement the classifier, we use the LeafRank algorithm with the same parameters as previously.

For each scoring function, we compute the empirical VUS as well as the local empirical VUS for the three proportions $u \in \{0.05, 0.1, 0.2\}$ averaged over the $V = 5$ validation set $\mathcal{D}^{(v)}$ and the $B = 5$ resample. We also compute the empirical standard deviations of the empirical VUS. The results are presented in Figure 6.4 and 6.5 through vizualisation plots and the exact values can be found in annex 6.5.

6.3.2 Discussion

Notice that in the cases where the number of classes equals 3, all the functions involved in the aggregation procedure are represented and that TR13 is used for the AggRR scoring function and the AggPFH scoring function. For the other dataset, we still present the same scoring functions as example but other bipartite scoring functions are involved in the aggregation procedures. The first interesting observation is that, in each of these experiments, Kendall aggregation clearly improves ranking accuracy, when measured in terms of VUS. More specifically, looking at the Unif2d and Unif2dSI datasets, we see clearly see the difference between a dataset with supports issue or without supports issue. The scoring function TR23 is outperformed by all the other pairwise functions. However the aggregated scoring function AggRR has a reasonable accuracy. Moreover, looking at the Cardiotocography results, all the aggregated scoring functions are much better that the bipartite scoring functions that they involve. For the psychometric data i.e. ERA, ESL, LEV and SWD (2-5) datasets, only the empirical VUS are reported since the number of classes is greater than 4. In these cases, it seems that all the aggregation procedures obtain comparable results and that the FH aggregation is better when the number of classes is low and that the RR decomposition is better when the number of classes is high (cf ERA dataset). Notice that for MQ2008, we clearly are in presence of the supports issue since only the scoring function TR23 is well below all the others. However, as in the case of Unif2dSI, the aggregated scoring function still obtain an accuracy very close to the best one, see table 6.5 in Annex. The second very interesting observation is that in every experiment the procedure TreeRank Tournament with the decomposition RR is the best of the TreeRank Tournament procedure. Even in the toy example with the supports issue Unif2dSI, the DTRRR outperforms all the other scoring functions. We explain this phenomenon by the fact that the TreeRank Tournament procedure is local, since at each step, the algorithm only uses the data available in the leaf to build the binary split. Finally, we see that all the scoring functions give comparable results except for the ERA dataset where the aggregation procedure outperforms the DTR procedure. However in these cases, the standard deviation are nearly at the same level of the VUS criterion so it is more difficult to state statiscal conclusions.

6.4 Comparison with competitors

In this section, we compare the algorithms proposed in the previous section with some competitors. We have selected 3 type of procedures, based respectively on the boosting principle RankBoost, the two other on the SVM heuristic (RankSVM and RankRLS). All these methods are based on the same principle, they solve a binary problem of classification over the pairs of observations.

6.4.1 Description of the competitors

Here, we describe the principle of the competitors as well as the parameters we use for each one.

The goal of the RankBoost algorithm, proposed in [Freund *et al.*, 2003], is to build a scoring function s minimizing the number of discordant pairs by combining weak predictors learned over a resample of the weighted learning dataset. At each iteration $t \in \{1, \dots, T\}$, the first step consists in building a weak predictor s_t from the training sample. The second step consists in giving a weight a_t for the scoring function s_t based of the prediction error. Finally, the third step we give weights to all the observations to create the next weighted learning dataset. At the end, the output scoring function is

$$s = \sum_{t=1}^{T} a_t s_t.$$

For the numerical experiments, we implement in matlab the RankBoost adapting to the ordinal data case the version of A. Rakotomamonjy implemented for the binary case. We choose to use as weak predictors binary scoring functions of the form

$$\forall x \in \mathcal{X},\ s_t(x) = \begin{cases} +1, \text{ if } x^{(j)} > \theta \\ -1, \text{ otherwise.} \end{cases}$$

For each experiment, the scoring function of the RankBoost procedure are implemented with $T = 30$ and we call it "RBpw".

The two other methods, RankSVM and RankRLS, are also minimizing the number of discordant pairs. Both of them are solving with an SVM type heuristic a criterion of the form :

$$\frac{2}{n(n-1)} \sum_{i<j} d(Y_i - Y_j, s(X_i) - s(X_j)) + \lambda \|s\|_{\mathcal{K}}$$

where $s(x) = \sum_{i=1}^{n} a_i K(x, X_i)$, $\lambda \in \mathbb{R}$ is the parameter of regularization and $\|.\|_{\mathcal{K}}$ is the RKHS-norm (Reproducing Kernel Hilbert Space) associated to the kernel $\mathcal{K}$ and $d(.,.)$ is a cost function.

For the RankSVM method, the cost function is the hinge loss i.e. $d(u, v) = max(1 - uv, 0)$ (see [Herbrich *et al.*, 2000]), and we use the implementation of T. Joachims available at http://svmlight.joachims.org/. We choose to parametrize the

algorithm with linear and Gaussian kernels with respective parameters $C = 20$ and $(C, \gamma) = (0.01, 1.0)$ and named them "SVMl" and "SVMg".

For the RankRLS method, the cost function is the quadratic loss i.e. $d(u, v) = (u - v)^2$ (see [Pahikkala *et al.*, 2007]), and we use the implementation available at http://www.tucs.fi/RLScore. We choose to parametrize the algorithm with linear and Gaussian kernels with respective parameters ("$bias = 1$") and ($\gamma = 0.01$) and named them "RLSl" and "RLSg". Notice that the selection of the intercept on a grid is performed through a leave-one-out procedure.

6.4.2 Results and discussion

We select the three following algorithms : Kendall Aggregation with the round robin decomposition, Kendall Aggregation with the FH decomposition and the DuelTreeRank procedure with the round robin decomposition. Indeed, they give the best performances overall the 11 experiments and we compare them to the competitors that we introduce in the previous subsection. We choose to present the results through visualization plots where the empirical VUS (and the local empirical VUS for the three proportions $u \in \{0.05, 0.1, 0.2\}$ when the number of classes equals 3) are represented as well as their empirical standard deviations.

Discussion.

The first notable thing is that the RankBoost procedure is always outperformed by the TreeRank procedures except for the Unif2dSI dataset. In this last case, it seems that the structure of the split, i.e. thresholding a coordinate, is the right thing to do. For the TreeRank methods the result can be much improved using LRCart as a LeafRank, i.e. the classifier used at each recursive step, instead of LRsvm (see [Robbiano, 2013]). As previously, all the procedures obtain very good results for the Gauss2d dataset. For the Cardiotocography dataset, the TreeRank procedure are much better than the others. We recall that for the feature space dimension equals 20 and that there is an important supports issue for this dataset that explains good performances of the TreeRank methods. More surprisingly, the linear kernels methods are better than the respective procedure using Gaussian kernel. For the Psychometric data set, all the procedures have very comparable accuracy, except in the case of ERA experiments where the kernel methods outperform all the others and SWD 2-5 where the TreeRank procedure are much better. Looking at the datasets, we can see that in the case of ERA, the feature space equals 4, the number of classes 7 or 9 and for the SWD 2-5 the feature space equals 10 and the number of classes is 4. In general, we can see that the TreeRank procedures are better when the feature space dimension is greater than the number of classes. Indeed, the principle of the TreeRank procedures are to estimate the ROC surface, a tool of dimension equal to the number of classes. If the feature space is equal or lower than the number of classes, estimating the ROC surface is more complicated than estimating the regression function. But, when the dimension of feature space is

much larger than the number of classes, the estimation of the ROC surface can be seen as a reduction of the dimension of the problem and is the right object to estimate. Despite the simplicity of this explanation, this explain most of the numerical results we obtain. Notice that for the MQ2008 dataset, only linear kernels are used because the procedures with Gaussian kernels are not feasible. In this case, all the performances in term of VUS look the same but the Kendall aggregation procedure outperforms the other methods for the localVUS. This is really important since in this application the goal is to recover the most pertinent web pages for each query.

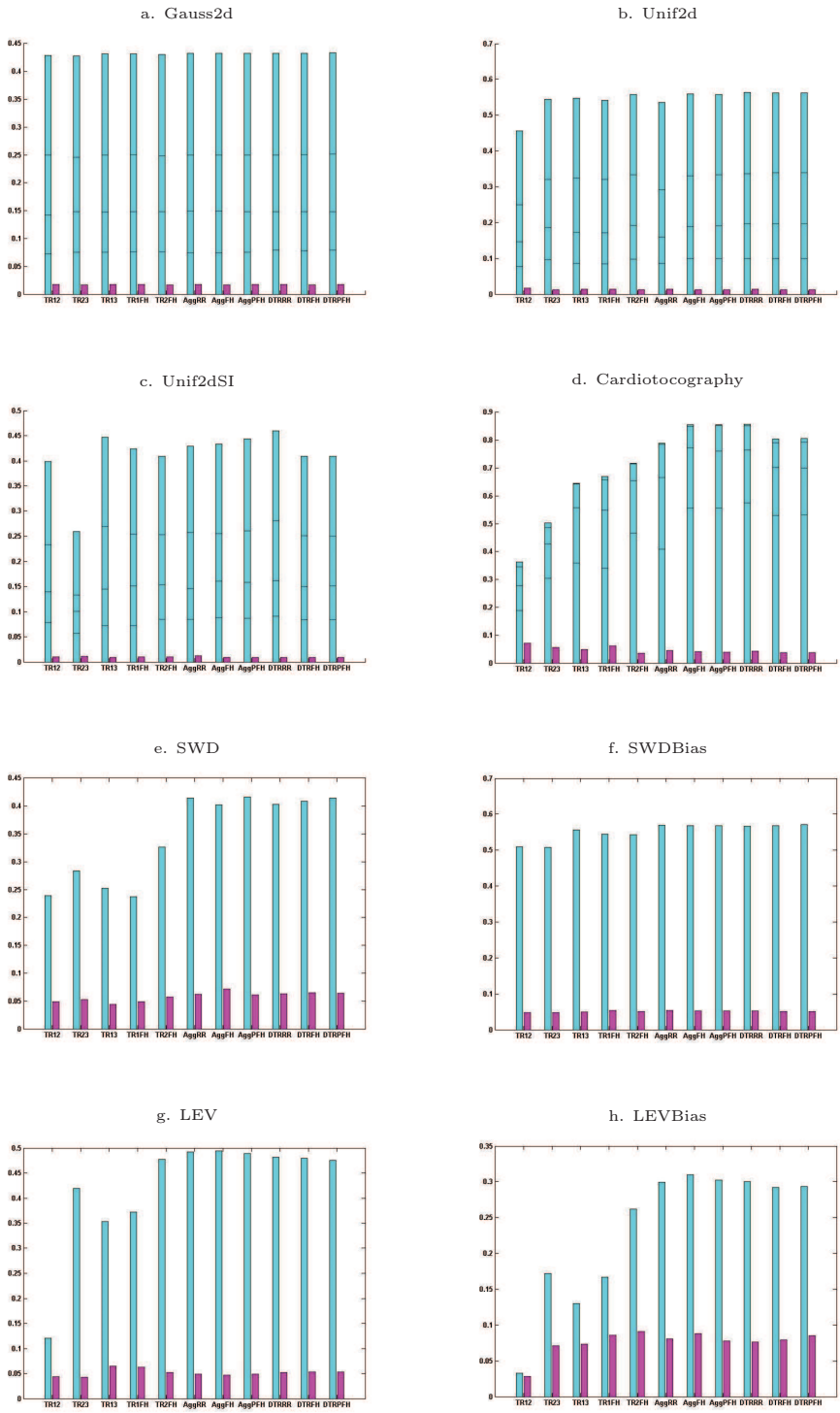


Figure 6.4: Comparison of 11 scoring functions : representation of the empirical VUS and the local VUS in cyan, and the standard deviation in mauve. From left to right : TR12, TR23, TR13, TR1FH, TR2FH, AggRR, AggFH, AggPFH,DTRRR, DTRFH, DTRPFH.

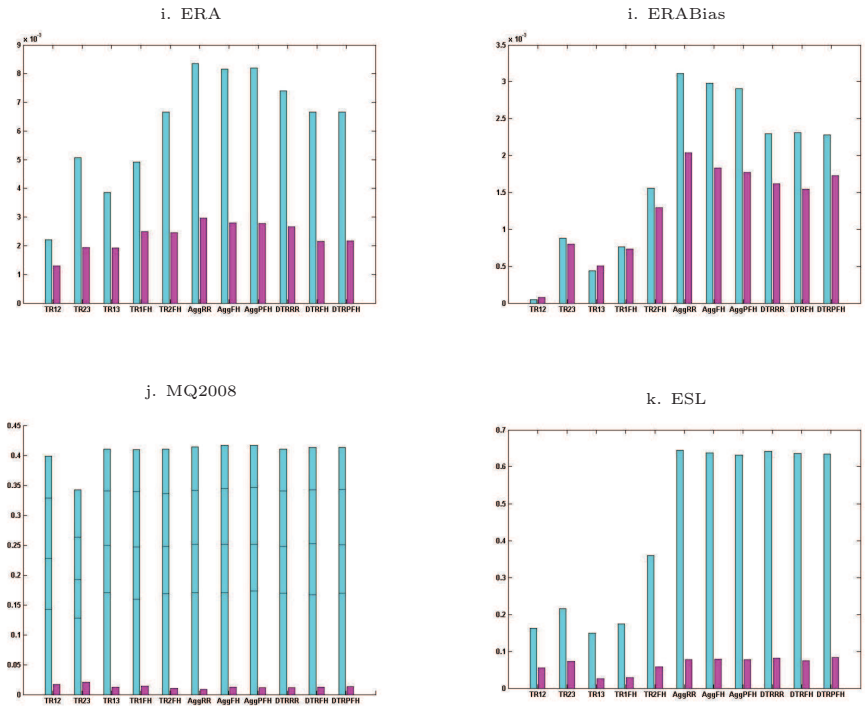


Figure 6.5: Comparison of 11 scoring functions : representation of the empirical VUS and the local VUS in cyan, and the standard deviation in mauve. From left to right : TR12, TR23, TR13, TR1FH, TR2FH, AggRR, AggFH, AggPFH,DTRRR, DTRFH, DTRPFH.

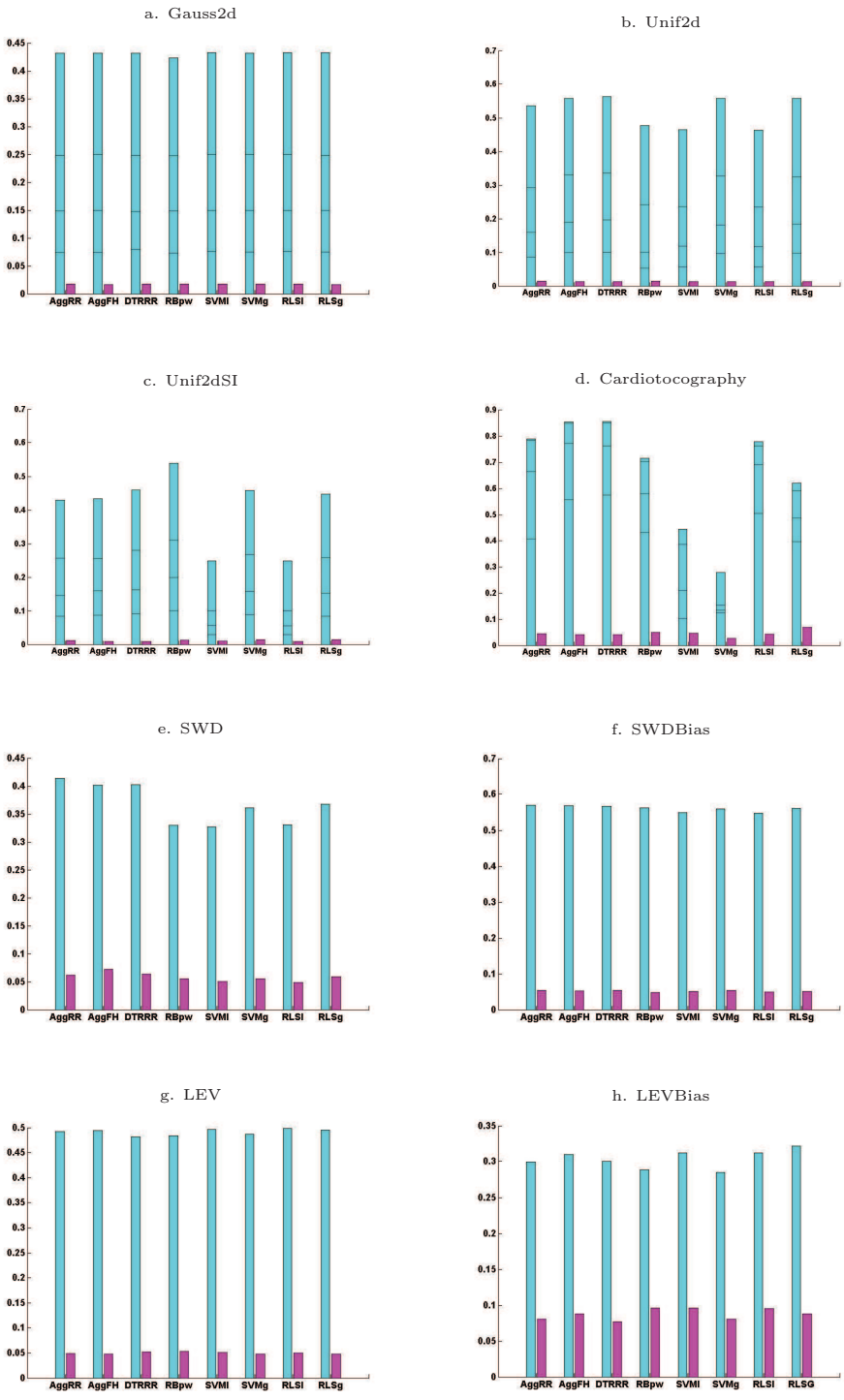


Figure 6.6: Comparison of 8 scoring functions : representation of the empirical VUS and the local VUS in cyan, and the standard deviation in mauve. From left to right : AggRR, AggFH, DTRRR, RBpw, SVMl, SVMg, RLSl, RLSg.

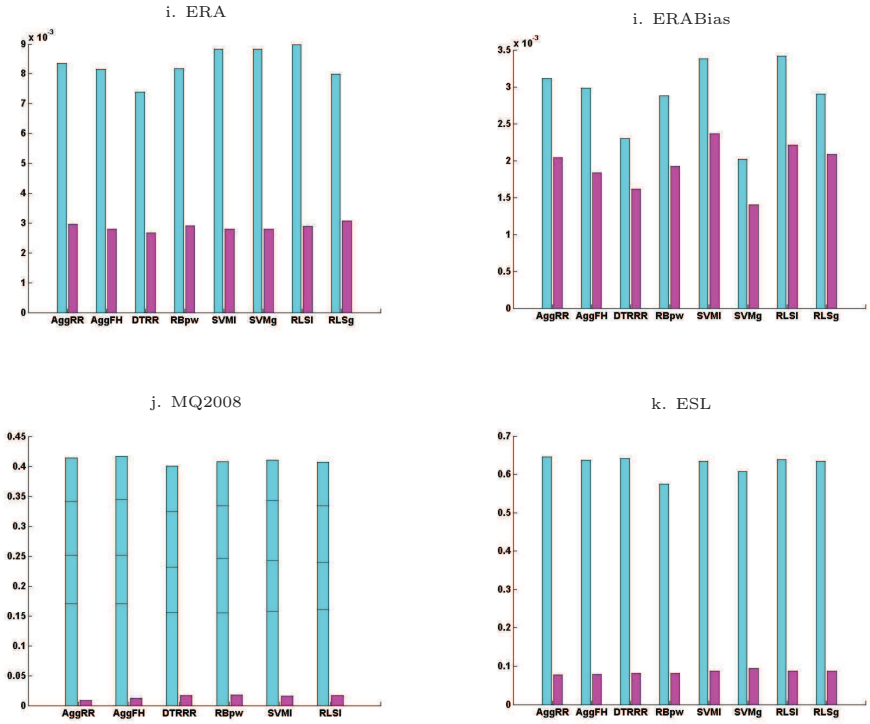


Figure 6.7: Comparison of 11 scoring functions : representation of the empirical VUS and the local VUS in cyan, and the standard deviation in mauve. From left to right : TR12, TR23, TR13, TR1FH, TR2FH, AggRR, AggFH, AggPFH,DTRRR, DTRFH, DTRPFH.

Table 6.5: Comparison of the VUS

Dataset	AggRR	AggFH	AggPFH	DTRRR	DTRFH	DTRPFH
Gauss2d	0.4327	0.4327	0.4328	0.4326	0.4326	0.4330
Unif2d	00.5359	0.5586	0.5585	0.5633	0.5627	0.5626
Unif2Dsi	0.4292	0.4334	0.4434	0.4592	0.4092	0.4089
Cardio	0.7897	0.8553	0.8559	0.8569	0.8052	0.8056
ERA 1-9	0.0031	0.0030	0.0029	0.0023	0.0023	0.0023
ERA 1-7	0.0083	0.0082	0.0082	0.0074	0.0067	0.0067
ESL 3-7	0.6454	0.6373	0.6310	0.6412	0.6356	0.6346
LEV 0-4	0.2993	0.3096	0.3023	0.3003	0.2922	0.2935
LEV 0-3	0.4924	0.4940	0.4886	0.4819	0.4790	0.4755
SWD 2-5	0.4144	0.4022	0.4168	0.4029	0.4090	0.4141
SWD 3-5	0.5695	0.5682	0.5680	0.5674	0.5685	0.5713
MQ2008	0.4150	0.4173	0.4175	0.4116	0.4137	0.4139

6.5 Annex - Numerical results

Table 6.6: Comparison of the VUS

Dataset	AggRR	AggFH	DTRRR	RBpw	SVMl	SVMg	RLSl	RLSg
Gauss2d	0.4327	0.4327	0.4326	0.4238	0.4334	0.4328	0.4337	0.4330
Unif2d	0.5359	0.5586	0.5633	0.4770	0.4644	0.5585	0.4641	0.5580
Unif2Dsi	0.4292	0.4334	0.4592	0.5393	0.2492	0.4581	0.2482	0.4479
Cardio	0.7897	0.8553	0.8569	0.7165	0.4450	0.2791	0.7788	0.6205
ERA 1-9	0.0031	0.0030	0.0023	0.0029	0.0034	0.0020	0.0034	0.0029
ERA 1-7	0.0083	0.0082	0.0074	0.0082	0.0088	0.0088	0.0090	0.0080
ESL 3-7	0.6454	0.6373	0.6412	0.5745	0.6337	0.6074	0.6387	0.6342
LEV 0-4	0.2993	0.3096	0.3003	0.2884	0.3124	0.2847	0.3122	0.3215
LEV 0-3	0.4924	0.4940	0.4819	0.4842	0.4968	0.4870	0.4983	0.4954
SWD 2-5	0.4144	0.4022	0.4029	0.3304	0.3278	0.3612	0.3316	0.3680
SWD 3-5	0.5695	0.5682	0.5674	0.5619	0.5493	0.5599	0.5483	0.5616
MQ2008	0.4150	0.4173	0.4010	0.4084	0.4113		0.4073	

Part III

Minimaxity and Ranking

Chapter 7

Minimax rates in bipartite ranking

The study of (minimax) learning rates in the context of classification/regression has been the subject of a good deal of attention in the machine-learning and statistical literature, see [Massart, 2000, Koltchinskii & Beznosova, 2005, Tsybakov, 2004, Audibert & A.Tsybakov, 2007, Lecué, 2008, Audibert, 2009, Srebro *et al.*, 2010] for instance. Under adequate smoothness/complexity assumptions on the regression function combined with a margin (or low noise) condition, minimax rates for the excess of misclassification risk have been proved in a variety of situations. Such analyses of best achievable rates of classification take into account the bias in the excess of misclassification risk and establish that plug-in classifiers (*i.e.* classifiers directly built from a nonparametric estimate of the regression function) may be optimal in the *minimax* sense.

In classification, when adding an assumption on the distribution of the regression function, rates faster than $n^{-1/2}$ and even faster than n^{-1} are achieved. The rates were obtained for plug-in classification rules in two papers. In [Audibert & A.Tsybakov, 2007], the authors estimate the regression function using the locally polynomial estimator. Moreover, the optimal rates are achieved without knowing the regularity and the margin parameters by aggregating the plug-in rules (see [Lecué, 2006]). More recently, the local multi-resolution estimation method (see [Monnier, 2012]) combined with the Lepski's method (see [Lepski *et al.*, 1997]), achieves the optimal and adaptive minimax rates. Both approaches firstly estimate the regression function and then threshold the estimated function at level 1/2.

In parallel, the *bipartite ranking*, akin to binary classification in the sense that it involves exactly the same probabilistic setup but of very different nature (it is *global* and not *local*), has recently received much interest in the statistical learning community. A rigorous formulation of the goal of bipartite ranking is given in [Clémençon *et al.*, 2008], where it is cast in terms of minimization of a *pairwise classification error*, called the *ranking risk*. Minimization of this error measure can be shown as equivalent to maximization of the so-termed "AUC criterion" [Hanley & McNeil, 1982], a widely used ranking criterion in practice. In the latter paper, a *low noise* assumption has been proposed, under which *Empirical Risk Minimization* (ERM) is shown to yield rates close to n^{-1}, under the restrictive assumption that an optimal ranking rule belongs to the set of candidates over which ERM is performed (*i.e.* assuming zero bias for the ranking method considered). In

[Clémençon & Vayatis, 2009a], plug-in ranking rules based on partitions (grids) of the input space have been considered in a less specific framework (relaxing the "zero bias" assumption namely), and have been proved to achieve rates slower than $n^{-1/2}$.

It is the major purpose of this chapter to pursue this analysis by considering more general *low noise* conditions together with smoothness/complexity assumptions for the regression function and study the rates attained by plug-in ranking rules, providing thus upper bounds for the *minimax rate of the expected excess of ranking risk*. We also use the aggregation with exponential weights in the bipartite ranking framework in order to obtain a method that can be adaptive to the parameters. The main result is that this procedure satisfies an oracle inequality. Then we study the impact of this inequality in two settings, one with the mild density assumption over the marginal of the observation and the other with the strong assumption (see [Audibert & A.Tsybakov, 2007]). When adding assumptions on the regression function, we obtain a new adaptive upper bound in the case of the mild density assumption. Moreover, when aggregating the plug-in estimators using the estimator of the regression function from [Audibert & A.Tsybakov, 2007], the procedure is adaptive to the parameters of the class of distributions under the strong density assumption. We also extend the optimality of the procedure by proving a minimax lower bound in dimension d.

The rest of the chapter is organized as follows. In section 7.1, we explain the notations and the bipartite ranking task. We define the ranking risk and a convexification of it using the hinge loss. Several margin assumptions are presented and equivalence links are stated. Preliminary results, based on the low noise conditions and linking the accuracy of nonparametric estimators of the regression function to the ranking risk of the related plug-in ranking rules are stated in section 7.2. In section 7.3, we describe the aggregation estimator using the convexified ranking risk and we show the oracle inequalities satisfied by the procedure of aggregation. In section 7.4, we present two adaptive minimax upper bounds for the excess ranking risk using the aggregated estimator. Finally, we state a minimax lower bound under the strong density assumption. The proof are deferred in section 7.7.

7.1 Theoretical background

Here, we introduce the main assumptions involved in the formulation of the bipartite ranking problem and recall the important results which are used in the following analysis, giving an idea of the nature of the bipartite ranking problem.

7.1.1 Probabilistic setup and first notations

In this chapter, (X, Y) denotes a pair of random variables, taking its values in the product space $\mathcal{X} \times \{-1, +1\}$ where $\mathcal{X}$ is typically a subset of an Euclidean space of (very) large dimension $d \geq 1$, $\mathbb{R}^d$ say. The r.v. X a vector of features for predicting the binary label Y. Let $p = \mathbb{P}\{Y = +1\}$ be the rate of positive instances.

The joint distribution of (X, Y) is denoted by P, X's marginal distribution by μ and the posterior probability by $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$, $x \in \mathcal{X}$. For the sake of simplicity and with no loss of generality, we assume that $\mathcal{X}$ coincides with $\mu(dx)$'s support. Additionally, the r.v. $\eta(X)$ is supposed to be continuous w.r.t. the Lebesgue measure.

We note $\textsc{Im}(\Phi)$ the range of any mapping Φ. We also denote by $\mathcal{B}(x, r)$ the closed Euclidean ball in $\mathbb{R}^d$ centered in $x \in \mathbb{R}^d$ and of radius $r > 0$. For any multi-index $s = (s_1, \ldots, s_d) \in \mathbb{N}^d$ and any $x = (x_1, ..., x_d) \in \mathbb{R}^d$, we set $|s| = \sum_{i=1}^d s_i, s! = s_1! \ldots s_d!, x^s = x_1^{s_1} \ldots x_d^{s_d}$ and $\|x\| = (x_1^2 + ... + x_d^2)^{1/2}$. Let D^s denote the differential operator $D^s = \frac{\partial^{s_1 + \ldots + s_d}}{\partial x_1^{s_1} \ldots \partial x_d^{s_d}}$ and $\lfloor \beta \rfloor$ the largest integer that is strictly less than $\beta \in \mathbb{R}$. For any $x \in \mathbb{R}^d$ and any $\lfloor \beta \rfloor$-times continuously differentiable real-valued function g on $\mathbb{R}^d$, we denote by g_x its Taylor polynomial expansion of degree $\lfloor \beta \rfloor$ at point x,

$$g_x(x') = \sum_{|s| \leq \lfloor \beta \rfloor} \frac{(x' - x)^s}{s!} D^s g(x).$$

7.1.2 Bipartite ranking

The bipartite ranking task consists in learning how to order the observations according to the label Y. Specifically, from a sample $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ with distribution P, we want to learn a scoring function $s : \mathcal{X} \to \mathbb{R}$ such that the order induced by s is the same as the order induced by η. In this case, the observations with label "+1" should have large values whereas the observations with label "−1" should have small values.

Pairwise classification

However, this is a functional tool and for this reason, it is complex to optimize from a theoretical and a computational perspective. For this reason, several authors have reformulated this problem as a pairwise classification problem (see [Freund *et al.*, 2003, Agarwal *et al.*, 2005, Clémençon *et al.*, 2008]). In this setup, the goal is, given (X, Y) and (X', Y') two random couples with distribution P, to determine whether $Y > Y'$ or not. We call the predictors ranking rule, namely a (measurable) function $r : \mathcal{X}^2 \to \{-1, +1\}$ such that $r(x, x') = 1$ when x' is ranked higher than x: the more pertinent a ranking rule r, the smaller the probability that it incorrectly ranks two instances drawn independently at random. Formally, optimal ranking rules are those that minimize the *ranking risk*:

$$L(r) \stackrel{def}{=} \mathbb{P}\left\{r(X, X') \cdot (Y' - Y) < 0\right\}. \tag{7.1}$$

A ranking rule r is said *transitive* iff $\forall (x, x', x'') \in \mathcal{X}^3$: "$r(x, x') = +1$ and $r(x', x'') = +1$" $\Rightarrow$ "$r(x, x'') = +1$". Observe that, by standard quotient set arguments, one can see that transitive ranking rules are those induced by scoring functions: $r_s(x, x') = 2 \cdot \mathbb{I}\{s(x') \geq s(x)\} - 1$ with $s : \mathcal{X} \to \mathbb{R}$ measurable. With a slight abuse of notation, we set $L(r_s) = L(s)$ for ranking rules defined through a scoring function s.

Optimality

It is easy to see that an optimal ranking rule is

$$r^*(x, x') = 2 \cdot \mathbb{I}_{\{\eta(x') > \eta(x)\}} - 1 \tag{7.2}$$

defined thanks to the regression function η, see Example 1 in [Clémençon *et al.*, 2008] for further details. Additionally, it should be noticed that one may derive a closed analytical form for the *excess of ranking risk* $\mathcal{E}(r) = L(r) - L^*$, with $L^* = L(r^*)$. For clarity, we recall the following result.

Lemma 7.1.1. (Ranking risk excess - [Clémençon *et al.*, 2008]) *For any ranking rule r, we have:*

$$\mathcal{E}(r) = \mathbb{E}\left[\left|\eta(X) - \eta(X')\right| \mathbb{I}\{r(X, X')(\eta(X') - \eta(X)) < 0\}\right].$$

The accuracy of a ranking rule is here characterized by the excess of ranking risk $\mathcal{E}(r)$, the challenge from a statistical learning perspective being to build a ranking rule, based on a training sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of i.i.d. copies of the pair (X, Y), with asymptotically small excess of ranking risk for large n.

We highlight the fact that, using a basic conditioning argument, the minimum ranking risk L^* can be expressed as a function of $\eta(X)$'s Gini mean difference:

$$L^* = p(1-p) - \frac{1}{2}\mathbb{E}[|\eta(X) - \eta(X')|]. \tag{7.3}$$

Hence, in contrast to binary classification where it is well-known that the learning problem is all the easier when $\eta(X)$ is bounded away from $1/2$, in bipartite ranking, Eq. (7.3) roughly says that the more spread the r.v. $\eta(X)$, the easier the optimal ranking of $\mathcal{X}$'s elements.

A continuum of classification problems

In addition, we emphasize the fact that the optimal ranking rule $r^*(x, x')$ can be seen as a (nested) collection of optimal cost-sensitive classifiers: the binary rule $r^*(x, X) = 2 \cdot \mathbb{I}\{\eta(X) > \eta(x)\} - 1$, related to the (regression) level set $G_t^* = \{x' \in \mathcal{X} : \eta(x') > t\}$ with $t = \eta(x)$, is optimal when considering the cost-sensitive risk $\mathcal{R}_\omega(C) = 2(1-p)\omega \cdot \mathbb{P}\{C(X) = +1 \mid Y = -1\} + 2p(1-\omega) \cdot \mathbb{P}\{C(X) = -1 \mid Y = -1\}$ with cost $\omega = \eta(x)$, see Proposition 15 in [Clémençon & Vayatis, 2010] for instance. Hence, while binary classification only aims at recovering the single level set $G_{1/2}^*$, which is made easier when $\eta(X)$ is far from $1/2$ with large probability (see [Massart & Nédélec, 2006] or [Tsybakov, 2004]), the ranking task consists in finding the whole collection $\{G_t^* : \; t \in \mathrm{Im}(\eta(X))\}$. Though of disarming simplicity, this observation describes well the main barrier for extending fast-rate analysis to the ranking setup: indeed, the random variable $\eta(X)$ cannot be far with arbitrarily high probability from all elements of its range.

Convexification of the ranking risk

Form a practical angle, to optimize the ranking risk is a real difficulty because the involved loss is not convex. In the classification framework where convex surrogates are widely used for practical purpose, it has been also used for theoretical issues ([Bartlett *et al.*, 2006], [Zhang, 2004] and [Lecué, 2006] for instance). Here, we propose to convexify the pairwise loss and we use this loss in our aggregation procedure (see 7.3). Notice that minimization of convexified pairwise loss has been studied in [Clémençon *et al.*, 2008]. We call any measurable function $f : \mathcal{X} \times \mathcal{X}' \to [-1, 1]$ a decision rule and we set the random variable $Z = (Y - Y')/2$. With this notation, we now present the convexification of the ranking risk that we use in this chapter.

Definition 7.1.1. *(Hinge ranking risk) For any decision function f, the hinge ranking risk is defined by*

$$A(f) \stackrel{def}{=} \mathbb{E}\phi\left(-f(X, X') \cdot Z\right), \tag{7.4}$$

where $\phi(x) = \max(0, 1 + x)$.

Notice that a ranking rule is a specific kind of decision rule. The next proposition gives a justification to strategies based on the minimization of the hinge ranking risk in order to obtain accurate ranking rules.

Proposition 7.1.2. *The minimizer of the ranking risk r^* is a minimizer of the hinge ranking risk A. We call $A^* = A(r^*)$.*

As for the ranking risk, there exists a close analytical form for the hinge ranking risk. This is the purpose of the next proposition.

Lemma 7.1.3. (Hinge ranking risk excess) *For any decision rule $f : \mathcal{X} \times \mathcal{X} \to [-1, 1]$, we have:*

$$A(f) - A^* = \mathbb{E}\left[\left|\eta(X) - \eta(X')\right| \left|f(X, X') - f^*(X, X')\right|\right].$$

The specific use of this surrogate is not fortunate and is due to its linearity. Using this property, we see that, for any ranking rule $r : \mathcal{X} \times \mathcal{X} \to \{-1, 1\}$, we have:

$$A(r) - A^* = 2(L(r) - L^*). \tag{7.5}$$

By thresholding a decision function, we can obtain a ranking rule. More precisely, for any decision rule f, we set $r_f(x, x') = 2\mathbb{I}\{f(x, x') \geq 0\} - 1$. We now link the excess of hinge ranking risk of a decision function f with the excess of ranking risk of its associated ranking rule. Using this definition, one can easily show that, for any decision rules $f : \mathcal{X} \times \mathcal{X} \to [-1, 1]$, we have:

$$L(r_f) - L^* \leq A(f) - A^*. \tag{7.6}$$

Thus, the minimization of the excess of hinge ranking risk provides a reasonable alternative for the minimization of the excess of ranking risk.

Plug-in ranking functions

Given the form of the Bayes ranking rule $r^*(X, X')$, it is natural to consider *plug-in* ranking rules, that is to say ranking rules obtained by "plugging" a non-parametric estimator $\widehat{\eta}_n(x)$ of the regression function η, based on a data sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, instead of $\eta(x)$ into Eq. (7.2):

$$\widehat{r}_n(x, x') \stackrel{def}{=} r_{\widehat{\eta}_n}(x, x'), \quad (x, x') \in \mathcal{X}^2.$$

The performance of predictive rules built via the plug-in principle has been extensively studied in the classification/regression context, under mild assumptions on the behavior of $\eta(X)$ in the vicinity of $1/2$ (see the references in [Audibert & A.Tsybakov, 2007] for instance) and on η's smoothness in particular. Similarly in the ranking situation, since one obtains as immediate corollary of Lemma 7.1.1 that $\mathcal{E}(\widehat{r}_n)$ is bounded by $\mathbb{E}[|\widehat{\eta}_n(X) - \eta(X)|]$, one should investigate under which conditions nonparametric estimators $\widehat{\eta}_n$ lead to ranking rules with fast rates of convergence of $\mathcal{E}(\widehat{r}_n)$ as the training sample size n increases to infinity.

7.1.3 Additional assumptions

Optimal ranking rules can be defined as those having the best possible rate of convergence of $\mathcal{E}(\widehat{r}_n)$ towards 0, as $n \to +\infty$. Therefore, the latter naturally depends on (X, Y)'s distribution. Following the footsteps of [Audibert & A.Tsybakov, 2007], we embrace the *minimax* point of view, which consists in considering a specific class $\mathcal{P}$ of joint distributions P of (X, Y) and to declare $\widehat{r}_n$ optimal if it achieves the best minimax rate of convergence over this class:

$$\sup_{P \in \mathcal{P}} \mathbb{E}\left[\mathcal{E}(\widehat{r}_n)\right] \sim \inf_{r_n} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\mathcal{E}(r_n)\right] \text{ as } n \to \infty,$$

where the infimum is taken over all possible ranking rules r_n depending on $(X_1, Y_1), \ldots, (X_n, Y_n)$. In order to carry out such a study, mainly three types of hypotheses shall be used. Here, smoothness conditions related to the real-valued function $\eta : \mathcal{X} \subset \mathbb{R}^d \to (0, 1)$ together with regularity conditions on the marginal $\mu(dx)$ and assumptions that we shall interpret as "spreadness" conditions for $\eta(X)$'s distribution are stipulated.

Complexity assumption

In the plug-in approach, the goal is to link closeness of $\widehat{\eta}_n(x)$ to $\eta(x)$ to the rate at which $\mathcal{E}(\widehat{r}_n)$ vanishes. Complexity assumptions for the regression function (CAR) stipulating a certain degree of smoothness for η are thus quite tailored for such a study. Here, focus is on regression functions η that belong to the $(\beta, L, \mathbb{R}^d)$-Hölder class of functions, denoted $\Sigma(\beta, L, \mathbb{R}^d)$, with $\beta > 0$ and $0 < L < \infty$. The latter

is defined as the set of functions $g : \mathbb{R}^d \to \mathbb{R}$ that are $\lfloor \beta \rfloor$ times continuously differentiable and satisfy, for any x, x' in $\mathbb{R}^d$, the inequality

$$|g(x') - g_x(x')| \leq L\|x - x'\|^\beta.$$

Remark 7.1.1. (ALTERNATIVE ASSUMPTIONS.) *We point out that more general* CAR *assumptions could be considered (see [Dudley, 1999] for instance), involving metric entropies or combinatorial quantities such as the* VC *dimension, more adapted to the study of the performance of empirical risk minimizers, see Chapter 3 and 5 for instances. The analysis is here restricted to the Hölder assumption.*

Marginal density assumption. Let strictly positive constants c_0 and r_0 be fixed. Recall first that a Lebesgue measurable set $A \subset \mathbb{R}^d$ is said to be (c_0, r_0)-regular iff $\forall r \in]0, r_0[$, $\forall x \in A$:

$$\lambda(A \cap \mathcal{B}(x, r)) \geq c_0 \lambda(\mathcal{B}(x, r)),$$

where $\lambda(B)$ denotes the Lebesgue measure of any Borelian set $B \subset \mathbb{R}^d$. The following assumption on the marginal distribution μ will be used in the sequel. Fix constants $c_0, r_0 > 0$ and $0 < \mu_{min} < \mu_{max} < \infty$ and suppose that a compact set $C \subset \mathbb{R}^d$ is given. The *strong density assumption* is said to be satisfied if the marginal distribution $\mu(dx)$ is supported on a compact and (c_0, r_0)-regular set $A \subset C$ and has a density f (w.r.t. the Lebesgue measure) bounded away from zero and infinity on A: $\mu_{min} \leq f(x) \leq \mu_{max}$ if $x \in A$ and $\mu(x) = 0$ otherwise.

The *mild density assumption* is said to be satisfied if the marginal distribution $\mu(dx)$ is supported on a compact and (c_0, r_0)-regular set $A \subset C$ and has a density f (w.r.t. the Lebesgue measure) bounded away from infinity on A: $f(x) \leq \mu_{max}$ for all $x \in A$.

Global low noise assumption

Here, we introduce an additional assumption for the function η. In classification, to obtain rates faster than $n^{1/2}$, one has to assume that the regression function η satisfies a low noise assumption. In ranking, such assumption has been used in [Clémençon *et al.*, 2008] and in the chapter 4 in order to show Theorem 4.2.1. Let $\alpha \in [0, 1]$. The following condition describes the behavior of the r.v. $\eta(X)$.

Assumption $\mathbf{NA}(\alpha)$. We have: $\forall (t, x) \in [0, 1] \times \mathcal{X}$,

$$\mathbb{P}\left\{|\eta(X) - \eta(x)| \leq t\right\} \leq C \cdot t^\alpha, \tag{7.7}$$

for some constant $C < \infty$.

Condition (7.7) above is void for $\alpha = 0$ and more and more restrictive as α grows. It clearly echoes Tsybakov's noise condition, introduced in [Tsybakov, 2004], which boils down to (7.7) with $1/2$ instead of $\eta(x)$. Whereas Tsybakov's noise condition is related to the behavior of $\eta(X)$ near the level $1/2$, condition (7.7) implies global properties for $\eta(X)$'s distribution, as shown by the following result.

Lemma 7.1.4. (Low noise and continuity) *Let $\alpha \in]0,1]$. Suppose that assumption* $\mathbf{NA}(\alpha)$ *is fulfilled, $\eta(X)$'s distribution is then absolutely continuous w.r.t. the Lebesgue measure on $[0,1]$. In addition, in the case where $\alpha = 1$, the related density is bounded by $C/2$.*

We point out that another low noise assumption has been proposed in [Clémençon *et al.*, 2008] in the context of the study of the performance of empirical (ranking) risk minimizers. The latter may be formulated as follows.

Assumption $\mathbf{LN}(\alpha)$. There exists $C < \infty$ such that:

$$\forall x \in \mathcal{X}, \ \ \mathbb{E}[|\eta(x) - \eta(X)|^{-\alpha}] \leq C. \tag{7.8}$$

Under the hypothesis above, it has been proved that minimizers of an empirical version of the ranking risk (7.1) of the form of a U-statistic have an excess of risk of the order $O_{\mathbb{P}}((\log n/n)^{1/(2-\alpha)})$ when optimization is performed over classes of ranking functions of controlled complexity (VC major classes of finite VC dimension for instance), that contains an optimal ranking rule (assuming thus zero bias for the ERM method), see Proposition 5 and Corollary 6 in [Clémençon *et al.*, 2008]. The following result describes the connection between these assumptions.

Proposition 7.1.5. (Noise assumptions) *The following assertions hold true.*

(i) If $\eta(X)$ fulfills Assumption $\boldsymbol{LN(\alpha)}$ *for $\alpha \in [0,1]$ then Assumption* $\boldsymbol{NA(\alpha)}$ *holds.*

(ii) Conversely, if $\eta(X)$ satisfies Assumption $\boldsymbol{NA(\alpha)}$ *then Assumption* $\boldsymbol{LN(\alpha - \varepsilon)}$ *holds for all $\varepsilon > 0$.*

In contrast to what happens for Tsybakov's noise condition, where α can be very large, up to $+\infty$, recovering in the limit Massart's margin condition [Massart, 2000], Assumption $\mathbf{NA}(\alpha)$ can be fulfilled for $\alpha \leq 1$ solely. Indeed, as may be shown by a careful examination of Lemma 7.1.4's proof, bound (7.7) for $\alpha > 1$ implies that $F'(\eta(x)) = 0$, denoting by F the cdf of $\eta(X)$. Therefore, it is obvious that the (probability) density of the r.v. $\eta(X)$ cannot be zero on its whole range $\mathrm{Im}(\eta) = \{\eta(x),\ x \in \mathcal{X}\}$.

Assumption $\mathbf{MA}(\alpha)$. The distribution P verifies the margin assumption $\mathbf{MA}(\alpha)$ with parameter $0 \leq \alpha \leq 1$ if there exists $C < \infty$ such that:

$$\mathbb{E}\left[\left|\mathbb{I}\{r(X,X') \neq r^*(X,X')\}\right|\right] \leq C(L(r) - L^*)^{\alpha/(1+\alpha)}, \tag{7.9}$$

for all measurable ranking rules $r : \mathcal{X} \times \mathcal{X} \to \{-1, +1\}$.

Assumption $\mathbf{MAK}(\alpha)$. The distribution P verifies the margin assumption $\mathbf{MAK}(\alpha)$ with parameter $0 \leq \alpha \leq 1$ if there exists $C < \infty$ such that:

$$\mathbb{E}\left[\left|f(X,X') - r^*(X,X')\right|\right] \leq C(A(f) - A^*)^{\alpha/(1+\alpha)}, \tag{7.10}$$

for all measurable decision functions $f : \mathcal{X} \times \mathcal{X} \to [-1, +1]$.

These conditions are introduced to control the variance of $\mathbb{I}\{r(X, X') \neq (Y - Y')/2\} - \mathbb{I}\{r^*(X, X') \neq (Y - Y')/2\}$. In particular, we use this control to state the oracle inequality 7.3.2. This type of conditions have been studied in classification in order to obtain fast rates (see [Boucheron *et al.*, 2005] for further details).

Equipped with these notations, we state the link between these assumptions.

Proposition 7.1.6. *If $\eta(X)$ fulfills Assumption **NA**(α) for $\alpha \in [0, 1]$ then Assumption **MA**(α) and **MAK**(α) hold.*

The theoretical results of this chapter are always stated using the condition **NA**(α). This is why, we do not need the inverse statement. Since in classification, such conditions are equivalent, it may be the same in ranking.

In the context of binary classification, by combining the CAR assumptions described above and Tsybakov's noise condition, optimal rates of convergence have been obtained in [Audibert & A.Tsybakov, 2007] and adaptive optimal rates in [Lecué, 2008]. In particular, it has been shown that, with the additional assumption that $\mu(dx)$ satisfies the *mild density assumption*, the minimax rate of convergence is $n^{-\beta(1+\alpha)/(d+\beta(2+\alpha))}$ and may be thus faster than $n^{-1/2}$ or even very close to n^{-1}, depending on the values taken by the parameters α and β. With the additional assumption that $\mu(dx)$ satisfies the *strong density assumption*, the minimax rate of convergence is $n^{-\beta(1+\alpha)/(2\beta+d)}$ and may be thus faster than $n^{-1/2}$ or even than n^{-1}. We shall now attempt to determine whether similar results hold in the ranking setup.

7.2 Comparison inequalities

It is the purpose of this section to show how the low noise assumption **NA**(α) enables to link the accuracy of a nonparametric estimate of $\eta(x)$ in terms of L_q-approximation error to the excess of ranking risk of the related plug-in ranking rule. Here, $\bar{\eta}$ is a Borel function on $\mathbb{R}^d$ and $\bar{r}(x, x') = 2 \cdot \mathbb{I}\{\bar{\eta}(x) \geq \bar{\eta}(x')\} - 1$ denotes the corresponding (plug-in) ranking function. The following results improve upon the bound stated in [Clémençon & Vayatis, 2009a], see Corollary 9 therein.

Proposition 7.2.1. (Risk excess and L_q-error) *Let $\alpha \in]0, 1[$ and assume that Assumption **NA**(α) is fulfilled. Then, the excess of ranking risk can be bounded as follows: there exists a constant $C < \infty$, such that for any distribution P and all approximant $\bar{\eta}$, we have*

$$L(\bar{\eta}) - L^* \leq C \|\eta - \bar{\eta}\|_\infty^{1+\alpha}. \tag{7.11}$$

In addition, we have : $\mathbb{P}\{\bar{r}(X, X') \neq r^*(X, X')\} \leq C\|\eta - \bar{\eta}\|_\infty^\alpha$, where (X, X') denotes a pair of independent r.v.'s drawn from $\mu(dx)$.

Let $1 \leq q < \infty$. There exist finite constants $C_0(\alpha, q)$, $C_1(\alpha, q)$ such that, whatever the distribution P and the approximant $\bar{\eta}$:

$$L(\bar{\eta}) - L^* \leq C_0(\alpha, q) \|\eta - \bar{\eta}\|_q^{\frac{q(1+\alpha)}{q+\alpha}} \tag{7.12}$$

and $\mathbb{P}\{\bar{r}(X, X') \neq r^*(X, X')\} \leq C_1(\alpha, q) \|\eta - \bar{\eta}\|_q^{\frac{q}{q+\alpha}}$.

These inequalities permit to derive bounds for the expected excess of ranking risk of plug-in ranking rules directly (by taking the expectation). Considering $L_\infty(\mathbb{R}^d, \mu)$-error for instance, the existence of nonparametric *locally polynomial* estimators (LP) $\widehat{\eta}_n$, optimal in the minimax sense, such that

$$\sup_{\eta \in \Sigma(\beta, L, \mathbb{R}^d)} \mathbb{E}\left[\|\widehat{\eta}_n - \eta\|_\infty^m\right] \leq C \left(\log n / n\right)^{m\beta/(2\beta+d)}, \tag{7.13}$$

for any $m > 0$, has been shown in [Stone, 1982] under the strong density assumption. With $m = 1 + \alpha$, this bound combined with Eq. (7.11) leads to an upper bound of the order $(\log n/n)^{(1+\alpha)\beta/(2\beta+d)}$ for the maximum expected excess of ranking risk of the rule $\widehat{r}_n = r_{\widehat{\eta}_n}$. Upper bound results, related to the MSE based on the $L_2(\mathbb{R}^d, \mu)$-error measure, established in [Yang, 1999] (see also [Stone, 1982] in a more restrictive framework, stipulating that the strong density assumption is fulfilled) claim that there exist nonparametric estimators of the regression function that attain the minimax rate $n^{-2\beta/(2\beta+d)}$ uniformly over the class $\Sigma(\beta, L, \mathbb{R}^d)$, yielding an upper bound of the order $n^{-2\beta(1+\alpha)/((2\beta+d)(2+\alpha))}$ for the maximum expected excess of ranking risk of the corresponding plug-in ranking functions.

However, although the comparison inequalities stated above are useful from a technical perspective (refer to section 7.7), as will be shown in section 7.4, such bounds are not optimal: in the L_∞ case, an extra logarithm factor appears in the rate thus obtained and in the L_2 situation, the exponent involved in the rate is even suboptimal.

7.3 Oracle inequalities for the aggregation procedure

In this section, we describe how to aggregate ranking rules into an accurate decision rule for the hinge ranking risk. We propose a procedure that uses exponentials weights. This kind of procedure is very popular in machine learning and has been studied in many context such as regression (see [Rigollet & Tsybakov, 2011], [Dalalyan & Tsybakov, 2008] and [Alquier & Lounici, 2011]), aggregation of experts (see [Cesa-Bianchi & Lugosi, 2006] for instance) and classification (see [Lecué, 2006]). We show that the obtained decision rule satisfies an oracle inequality which can be used to achieve minimax upper bounds (see 7.4). The proof of the theorem is an adaptation to the ranking case of the one in [Lecué, 2006].

7.3.1 Aggregation via exponential weights

The ranking rules $r_1, \dots, r_M$ are given and the goal of the aggregation method is to mimic the performance of the best of them according to the excess risk and under the low noise assumption. We define the exponential aggregate decision rule as

$$\tilde{f}_n = \sum_{m=1}^{M} w_m^{(n)} r_m \tag{7.14}$$

where the weights $w_m^{(n)}$ are

$$w_m^{(n)} = \frac{\exp(\sum_{i \neq j} -Z_{ij} r_m(X_i, X_j))}{\sum_{k=1}^{M} \exp(\sum_{i \neq j} -Z_{ij} r_m(X_i, X_j))}, \forall m = 1, \dots M.$$

Notice that we call it $\tilde{f}_n$ because this function takes its values in $[-1, 1]$. The functions $r_1, \dots, r_M$ take their values in $\{1; -1\}$, we have,

$$w_m^{(n)} = \frac{\exp(-n(n-1)A_n(r_m))}{\sum_{k=1}^{M} \exp(-n(n-1)A_n(r_k))}, \forall m = 1, \dots M, \tag{7.15}$$

where $A_n(r_m) = \frac{1}{n(n-1)} \sum_{i \neq j} \max(0, 1 - Z_{ij} r_m(X_i, X_j))$ is the empirical hinge ranking risk of the ranking rule r_m. Using the equality (7.5), the weights can be rewritten in terms of the empirical risks of r_m's

$$w_m^{(n)} = \frac{\exp(-2n(n-1)2L_n(r_m))}{\sum_{k=1}^{M} \exp(-2n(n-1)L_n(r_k))}, \forall j = 1, \dots M,$$

We call this procedure aggregation with exponential weights (AEW). The idea behind this procedure is to give more weight to the ranking rules that have the smaller empirical performance in order to mimic the accuracy of the empirical (hinge ranking) risk minimizer (ERM). The next result states that the AEW has similar performance as the ERM estimator up to a $(\log M)/n$ term.

Proposition 7.3.1. *Let $M \geq 2$ be an integer, $f_1, \dots, f_M$ be M decision rules on $\mathcal{X} \times \mathcal{X}$. For any $n \in \mathbb{N}^*$, the aggregate $\tilde{f}_n$ estimator defined in 7.14 with weights 7.15 satisfies*

$$A_n(\tilde{f}_n) \leq \min_{j=1,\dots,M} A_n(f_j) + \frac{\log M}{n}.$$

The main benefits of the AEW procedure are that it does not need a minimization algorithm and is less sensitive to overfitting because the output decision rule is a mixture of several ranking rules whereas ERM only involves one ranking rule.

7.3.2 An oracle inequality

We now provide the main tool of this chapter, an oracle inequality for the excess of hinge ranking risk. The goal of an oracle inequality is to show that an estimator is nearly as good as the best one of a given collection (see [Massart, 2006] for example

in model selection). Here, the goal of this oracle inequality is to state that the procedure AEW 7.14 has asymptotically the same performance as the best one among the convex hull formed by a finite set of decision functions.

Theorem 7.3.2. *(Oracle inequality) Let $\alpha \in (0,1]$. We assume that NA(α) holds. We denote by $\mathcal{C}$ the convex hull of a finite set $\mathcal{F}$ of functions $f_1, \ldots, f_M$ with values in $[-1,1]$. Let $\tilde{f}_n$ be the aggregate estimator introduced in (7.14). Then, for any integers $M \geq 3, n \geq 1$ and any $a > 0$, $\tilde{f}_n$ satisfies the inequality*

$$\mathbb{E}[A(\tilde{f}_n) - A^*] \leq (1+a) \min_{f \in \mathcal{C}} (A(f) - A^*) + C \left(\frac{\log M}{n} \right)^{\frac{\alpha+1}{\alpha+2}},$$

where $C > 0$ is a constant depending only on a.

In [Lecué, 2006], the author shows that the rate $\left(\frac{\log M}{n}\right)^{\frac{\alpha+1}{\alpha+2}}$ is optimal in a minimax sense. For the moment, we do not have such result of optimality, however, the rate in the oracle inequality is the same. Using this tool, we can state an oracle inequality for the excess of ranking risk.

Corollary 7.3.3. *(Oracle inequality for the ranking risk) Let $\alpha \in (0,1]$, $M \geq 3$ and $\{r_1, \ldots, r_M\}$ be a finite set of prediction rules. We assume that NA(α) holds. Let $\tilde{f}_n$ be the aggregate estimator introduced in 7.14. Then, for any integers $M \geq 3, n \geq 1$ and any $a > 0$, $\tilde{f}_n$ satisfies the inequality*

$$\mathbb{E}[L(r_{\tilde{f}_n}) - L^*] \leq 2(1+a) \min_{f \in \mathcal{C}} (L(r_f) - L^*) + C \left(\frac{\log M}{n} \right)^{\frac{\alpha+1}{\alpha+2}},$$

where $C > 0$ is a constant depending only on a and the constant in the condition NA(α).

Proof. Using inequalities 7.5 and 7.6 combine with Theorem 7.3.2, we immediately get the desired result. □

This oracle is the main tool to obtain the minimax rates in Theorems 7.4.1, 7.4.2 and 7.4.4 using an estimator based on the AEW procedure.

7.4 Minimax rates

Here, we present the adaptive minimax upper bounds in bipartite ranking in two cases, specifically under the mild assumption and the strong assumption. The estimators of the regression function used are the same as in classification (see [Lecué, 2006] and [Audibert & A.Tsybakov, 2007]).

7.4.1 The "mild" case

In this section, we assume that the regression function η belongs to a Hölder class of functions. An important result from [Kolmogorov & Tikhomirov, 1961], on the complexity of Hölder classes, says that :

$$\mathcal{N}\left(\Sigma(\beta, L, [0,1]^d), \varepsilon, L^\infty([0,1]^d)\right) \leq C\varepsilon^{-\frac{d}{\beta}}, \forall \varepsilon > 0$$

where the left hand side is the ε-entropy of the $(\beta, L, [0,1]^d)$-Hölder class w.r.t. to the $L^\infty([0,1]^d)$ norm and C is a constant depending only on β and d. We now introduce the first class of distributions for the random couple (X, Y).

Definition 7.4.1. *Let $\alpha \leq 1$, β and L be strictly positive constants. The collection of probability distributions $P(dx, dy)$ such that*

1. *the marginal μ satisfies the mild density assumption with μ_{max},*
2. *the global noise assumption **NA(α)** holds,*
3. *the regression function belongs to Hölder class $\Sigma(\beta, L, \mathbb{R}^d)$,*

is denoted by $\mathcal{P}_{\alpha,\beta,\mu_{max}}$ (omitting to index it by the constant involved in the noise assumption for notational simplicity).

Let $\alpha \leq 1, \beta > 0$. For $\varepsilon > 0$, $\Lambda_\varepsilon(\beta)$ is an ε-net on $\Sigma(\beta, L, [0;1]^d)$ for the L^∞-norm, such that $\ln(\text{Card}(\Lambda_\varepsilon(\beta))) \leq C\varepsilon^{-d/\beta}$. We use the procedure 7.14 over the net $\Lambda_\varepsilon(\beta)$ to define the estimator :

$$\tilde{f}_n^{\varepsilon,\beta} = \sum_{g \in \Lambda_\varepsilon(\beta)} w^n(r_g) r_g \tag{7.16}$$

where $r_g(x, x') = 2\mathbb{I}\{g(x) > g(x')\} - 1$ and we call the associated ranking rule $\tilde{r}_n^{\varepsilon,\beta}$. This is an adaptation to the ranking case of the estimator in [Lecué, 2006]. We now state the minimax upper bound for the excess of ranking risk over the class of distribution with the mild assumption.

Theorem 7.4.1. (Upper bound: mild case) *There exists a constant $C > 0$ such that for all $n \geq 1$, the maximum expected excess of ranking risk of the aggregation rule defines in 7.16 $\varepsilon_n = n^{-\alpha\beta/(d+\beta(2+\alpha))}$, is bounded as follows:*

$$\sup_{P \in \mathcal{P}_{\alpha,\beta,\mu_{max}}} \mathcal{E}(\tilde{r}_n^{\varepsilon,\beta}) \leq C \cdot n^{-\frac{\beta(1+\alpha)}{d+\beta(2+\alpha)}}. \tag{7.17}$$

where C depends on d, β and α.

To obtain an estimator adaptive to the smoothness and the margin coefficients, we aggregate classifiers $\tilde{r}_n^{(\varepsilon,\beta)}$ for (ε, β) in a finite grid. We split the sample in two sets, the first set $D_m^{(1)}$ of size $m = n - \lfloor n/\ln n \rfloor$ is used to build the plug-in classifiers

and the second one $D_l^{(2)}$ of size $l = \lfloor n/\ln n \rfloor$ to obtain the weights. We define the grid $\mathcal{G}$ of values for (ε, β):

$$\mathcal{G} = \left\{ (\varepsilon_k, \beta_p) = \left(m^{-\phi_k}, \frac{p}{\ln n} \right) | \phi_k = \frac{k}{\ln n}, k \in \{1, \dots, \lfloor \ln(n)/2 \rfloor\}, p \in \left\{ 1, \dots, \lfloor \ln(n) \rfloor^2 \right\} \right\}.$$

We propose the ranking rule $\tilde{r}_n^{adp}$ which is the sign of the decision function

$$\tilde{f}_n^{adp} = \sum_{(\varepsilon,\beta)\in\mathcal{G}} w^{(l)}(r_m^{\varepsilon,\beta}) r_m^{\varepsilon,\beta}$$

where the weights $w^{(l)}(r)$ are those defined in 7.15 using the dataset $D_l^{(2)}$ and $r_m^{\varepsilon,\beta}$ is the ranking rule associated to the decision function introduced in equation 7.16 using the dataset $D_m^{(1)}$.

Theorem 7.4.2. (Adaptivity in α and β) *Let K be a compact subset of $]0;1[\times]0;\infty[$. There exists a constant $C > 0$ such that for all $n \geq 1$, for any $(\alpha, \beta) \in K$, the maximum expected excess of ranking risk of the estimator $\tilde{r}_n^{adp}$ is bounded as follows:*

$$\sup_{P\in\mathcal{P}_{\alpha,\beta,\mu_{max}}} \mathcal{E}(\tilde{r}_n^{adp}) \leq C \cdot n^{-\frac{\beta(1+\alpha)}{d+(2+\alpha)\beta}}. \tag{7.18}$$

The cardinality of Σ_{ε_n} is an exponential of n so the estimators $\tilde{r}^{\varepsilon_n,\beta}$, for a given (ε_n, β), are not easily implementable. However the procedure is very interesting from a theoretical standpoint since it is adaptive to the parameters and it achieves fast rates when $\alpha\beta > d$. Finally, notice that this estimator can achieve fast rates when $\alpha\beta > d$ and close to $n^{-2/3}$ when the regression function is very smooth.

7.4.2 The "strong" case

Now, we introduce the second case, namely the strong density assumption. The class of distributions is given in the next definition.

Definition 7.4.2. *Let $\alpha \leq 1$, β and L be strictly positive constants. The collection of distributions probabilities $P(dx, dy)$ such that*

1. *the marginal μ satisfies the strong density assumption with μ_{max} and with μ_{min},*
2. *the global noise assumption **NA(α)** holds,*
3. *the regression function belongs to Hölder class $\Sigma(\beta, L, \mathbb{R}^d)$,*

is denoted by $\mathcal{P}_{\alpha,\beta,\mu_{max},\mu_{min}}$ (omitting to index it by the constant involved in the noise assumption assumption for notational simplicity).

An upper bound for the minimax rate is proved by exhibiting a sequence of ranking rules attaining the latter. Here we consider the same estimator as that studied in [Audibert & A.Tsybakov, 2007] (see section 3 therein). Let $K : \mathbb{R}^d \to \mathbb{R}$ be a Parzen-Rosenblatt kernel such that K is bounded away from 0 on a neighborhood of 0 in $\mathbb{R}^d$, $\int(1+||x||^{4\beta})K^2(x)dx < \infty$ and $\sup_x(1+||x||^{2\beta})K^2(x) < \infty$. Fix $l \in \mathbb{N}$ and a bandwidth $h > 0$, set $U(u) = (u^s)_{|s|\leq l}$, $Q = (Q_{s_1,s_2})_{|s_1|,\,|s_2|\leq\lfloor l\rfloor}$ with $Q_{s_1,s_2} = \sum_{i=1}^n (X_i - x)^{s_1+s_2}K((X_i-x)/h)$ and $B_n = (B_{s_1,s_2})_{|s_1|,\,|s_2|\leq\lfloor\beta\rfloor}$ with $B_{s_1,s_2} = (nh^d)^{-1}\sum_{i=1}^n((X_i-x)/h)^{s_1+s_2}K((X_i-x)/h)$. Consider the estimator $\widehat{\eta}_{n,h}(x)$ equal to the locally polynomial estimate

$$\widehat{\eta}^{LP}_{n,h}(x) = \sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right) U^t(0)Q^{-1}U(X_i - x)$$

when $\widehat{\eta}^{LP}_{n,h}(x) \in [0,1]$ and B_n's smallest eigenvalue is larger than $1/\log n$, and to 0 otherwise.

Theorem 7.4.3. (A MINIMAX UPPER BOUND) *There exists a constant $C > 0$ such that for all $n \geq 1$, the maximum expected excess of ranking risk of the plug-in rule $\widehat{r}_n(x,x') = 2\cdot\mathbb{I}\{\widehat{\eta}^{LP}_{n,h_n}(x') > \widehat{\eta}^{LP}_{n,h_n}(x)\} - 1$, with $h_n = n^{-1/(2\beta+d)}$ and $l = \lfloor\beta\rfloor$, is bounded as follows:*

$$\sup_{P\in\mathcal{P}_{\alpha,\beta,L}} \mathcal{E}(\widehat{r}_n) \leq C\cdot n^{-\frac{\beta(1+\alpha)}{d+2\beta}}. \tag{7.19}$$

The plug-in estimator defined in the last theorem depends only on β. To obtain an estimator adaptive to the smoothness coefficient, we aggregate classifiers $\widehat{r}_n^{(\beta)}$ for β in a finite grid. We split the sample in two sets, the first set of size $m = n - \lfloor n/\ln n\rfloor$ is used to build the plug-in classifiers and the second one of size $l = \lfloor n/\ln n\rfloor$to obtain the weights. We define the set $\mathcal{F}$ of of plug-in classifiers using the training sample $D_m^1 = (X_i, Y_i)_{1\leq i\leq m}$:

$$\mathcal{F} = \left\{\widehat{r}_n^{(\beta_k)} | \beta_k = \frac{kd}{\ln(n) - 2k}, k \in \{1,\dots,\lfloor\ln(n)/2\rfloor\}\right\}.$$

Using the validation sample $D_l^2 = (X_i, Y_i)_{m+1\leq i\leq n}$, we build the weights, for all $r \in \mathcal{F}$

$$w_n^{(l)}(r) = \frac{\exp(\sum_{i=m+1}^n Y_i r(X_i))}{\sum_{\bar{r}\in\mathcal{F}}\exp(\sum_{i=m+1}^n Y_i\bar{r}(X_i))}$$

Finally, our ranking rule is $\widehat{r}^{adp} = \mathrm{sign}(\widehat{\mathrm{f}}^{\mathrm{adp}})$, where $\widehat{f}^{adp} = \sum_{r\in\mathcal{F}} w_n^{(l)}(r)r$.

Theorem 7.4.4. (ADAPTIVITY IN β) *Let K be a compact subset of $]0;1[\times]0;\infty[$. There exists a constant $C > 0$ such that for all $n \geq 1$, for any $(\alpha,\beta) \in K$ such that $\alpha\beta \leq d$, the maximum expected excess of ranking risk of the estimator $\tilde{r}^{adp}$ is bounded as follows:*

$$\sup_{P\in\mathcal{P}_{\alpha,\beta,\mu_{max},\mu_{min}}} \mathcal{E}(\widehat{r}^{adp}) \leq C\cdot n^{-\frac{\beta(1+\alpha)}{d+2\beta}}. \tag{7.20}$$

Remark 7.4.1. (FAST, BUT NOT SUPER-FAST, RATES.) *Notice that, since $\alpha \leq 1$ rates faster than n^{-1} cannot be achieved by the plug-in rule $\widehat{r}_n$ defined in the theorem above, in spite of the optimality of the related estimator $\widehat{\eta}^{LP}_{n,h_n}$. However, for any $\alpha \in]0,1]$, fast rates can be attained (i.e. rates faster than $n^{-1/2}$), provided that the regression function is sufficiently smooth, when $\beta > d/2\alpha$ namely.*

7.5 A lower bound

For completeness, we now state a lower bound for the minimax rate of the expected excess of ranking risk in the strong density case. It holds in a specific situation, namely when $\alpha\beta \leq 1$. When $d = 1$, the result can be found in [Clémençon & Robbiano, 2011].

Theorem 7.5.1. (A MINIMAX LOWER BOUND) *Let $(\alpha, \beta) \in]0,1] \times \mathbb{R}_+^*$ such that $\alpha\beta \leq 1$. There exists a constant $C > 0$ such that, for any ranking rule r_n based on n independent copies of the pair (X, Y), we have: $\forall n \geq 1$,*

$$\sup_{P \in \mathrm{Pr}_{\alpha,\beta,\mu_{max},\mu_{min}}} \mathcal{E}(r_n) \geq C \cdot n^{-\frac{\beta(1+\alpha)}{d+2\beta}}.$$

When $d \geq 2$ the rate of convergence is always slower than $n^{-1/2}$. That means that we are not able to prove optimal fast rates for the excess ranking risk. In classification, the limitation is $\alpha\beta \leq d$, so optimal fast rates can be achieved in this situation (but not hyper fast).

For the mild case and the oracle inequality, mimicking the proof of theorem 7.5.1 does not give the same rates as the upper bounds. An explanation of the difficulties as well as the rates one can obtain are given in Appendix 7.8

7.6 Conclusion

The need for understanding the originality/specificity of bipartite ranking in regards to the (minimax) learning rates that can be attained, in comparison to classification rates in particular, motivates the present chapter. A global low noise assumption, extending the Mammen-Tsybakov condition originally proposed in the context of binary classification, is introduced, under which novel comparison inequalities, linking approximation error of a regression estimate and ranking risk of the corresponding plug-in rule, are proved. We investigate the aggregation with exponential weights of ranking rules. We state an oracle inequality for the aggregation procedure under a low noise assumption that achieves the same rate as in classification. This is the crucial point to obtain the adaptive upper bounds for the excess of ranking risk. In the mild density case, we establish a new upper bound that it is adaptive to the margin and the regularity parameters, with the same rates as in classification. By considering a specific (locally polynomial) regression estimator, we highlighted

the fact that fast rates can be achieved (by plug-in ranking rules in particular) in certain situations. We aggregate plug-in classifiers in order to obtain minimax adaptive rates of convergence, under a restrictive assumption over the parameters for all dimensions. A preliminary lower bound result showed that these rates are actually optimal in a restrictive situation. Moreover, in dimension 1, the aggregation procedure attains minimax adaptive fast rates. To the best of our knowledge, the present analysis, destined to be completed in regards to minimax lower bounds and adaptivity of the nonparametric estimators considered, is the first to state results of this nature.

7.7 Proofs

Proof of Proposition 7.1.2

Proof.

$$\begin{aligned} A(f) &= \mathbb{E}[1 - (f(X, X') \cdot Z)] \\ &= 1 - \mathbb{E}[f(X, X')(\eta(X)(1 - \eta(X'))) - f(X, X')(\eta(X')(1 - \eta(X)))] \\ &= 1 - \mathbb{E}[f(X, X')(\eta(X) - \eta(X'))] \end{aligned}$$

Finally to minimize A, we have to take $f^*(x, x') = 1$ when $\eta(x) \geq \eta(x')$ and $f^*(x, x') = -1$ otherwize. □

Proof of Lemma 7.1.3

Proof. Because f takes value in $[-1, 1]$

$$\begin{aligned} A(f) - A^* &= \mathbb{E}\left[-f(X, X') \cdot Z + f^*(X, X') \cdot Z\right] \\ &\quad \mathbb{E}\left[-f(X, X')\eta(X) + f(X, X')\eta(X') + f^*(X, X')\eta(X) - f^*(X, X')\eta(X')\right] \\ &\quad \mathbb{E}\left[(f(X, X') - f^*(X, X'))(\eta(X') - \eta(X))\right] \end{aligned}$$

Because f takes its values in $[-1, 1]$ and by definition of $f^*(X, X')$, we get the desired result. □

Proof of Lemma 7.1.4.

Let F denote $\eta(X)$'s cumulative distribution function. The first part of the lemma immediately results from the fact that $\mathbf{NA}(\alpha)$ can be rewritten as follows: $\forall (t, x) \in \mathbb{R}_+ \times \mathcal{X}$,

$$F(\eta(x) + t) - F(\eta(x) - t) \leq C \cdot t^{\alpha}.$$

The cdf F is thus absolutely continuous. Denote by ϕ the related density. Observe that, when $\alpha = 1$, the bound above can be written as $(F(\eta(x)+t)-F(\eta(x)-t))/t \leq C$. Letting then t tend to zero, one obtains that, for all $x \in supp(\mu)$, $2\phi(\eta(x)) \leq C$.

Proof of Proposition 7.1.5.

Hölder inequality combined with condition $\mathbf{NA}(\alpha)$ shows that $\mathbb{E}[\mathbb{I}\{|\eta(X)-\eta(x)|<t\}]$ is bounded by

$$c^{1/(1+\alpha)}\mathbb{E}[\mathbb{I}\{|\eta(X)-\eta(x)|<t\}\cdot|\eta(X)-\eta(x)|]^{\alpha/(1+\alpha)},$$

which quantity is clearly less than $c^{1/(1+\alpha)}t^{\alpha/(1+\alpha)}$. This permit to prove assertion (i).

Let $x\in\mathcal{X}$ and $\varepsilon>0$ be fixed. We have $\mathbb{E}[|\eta(x')-\eta(X)|^{-\alpha+\varepsilon}] = \int_0^{+\infty}\frac{\alpha}{t^{1+\alpha-\varepsilon}}\mathbb{P}\left\{|\eta(X)-\eta(x)|<t\right\}dt$. Using $\mathbf{NA}(\alpha)$ when integrating over $[0,1]$ and bounding simply the probability by 1 otherwise, this permits to establish assertion (ii).

Proof of Proposition 7.1.6

Proof. Recall that

$$(L(r)-L^*)=\mathbb{E}\left[\left|\eta(X)-\eta(X')\right|\mathbb{I}\{r(X,X')(\eta(X')-\eta(X))<0\}\right]$$

Minoring $\eta(X)-\eta(X')$ by t we obtain the lower bound

$$t\mathbb{E}\mathbb{I}\{r(X,X')(\eta(X')-\eta(X))<0\}\mathbb{I}\{\eta(X')-\eta(X)>t\}$$

which is greater (using the noise assumption) than

$$t\mathbb{E}[\mathbb{I}\{r(X,X')\neq r^*(X,X')\}]-Ct^{1+\alpha}$$

Optimizing in the parameter t, we obtain (for $t_0=\left(\frac{\mathbb{E}\mathbb{I}\{r(X,X')\neq r^*(X,X')\}}{C(1+\alpha)}\right)^{1/\alpha}$):

$$\mathbb{E}[\mathbb{I}\{r(X,X')\neq r^*(X,X')\}]\leq\frac{C(1+\alpha)}{C\alpha^{\alpha/(1+\alpha)}}(L(r)-L^*)^{\alpha/(1+\alpha)}$$

□

Proof of Lemma 7.2.1.

Lemma 7.1.1 yields

$$\mathcal{E}(r_{\bar\eta})=\mathbb{E}[|\eta(X)-\eta(X')|\mathbb{I}\{(X,X')\in\Gamma_{\bar\eta}\}],$$

where $\Gamma_{\bar\eta}=\{(x,x'):\ (\bar\eta(x')-\bar\eta(x))(\eta(x')-\eta(x))<0\}$. Observe that on the event $\Gamma_{\bar\eta}$, we have

$$\begin{aligned}|\eta(X)-\eta(X')| &\leq |\eta(X)-\bar\eta(X)|+|\eta(X')-\bar\eta(X')|\\ &\leq 2\ \|\eta-\bar\eta\|_\infty.\end{aligned}$$

Using now condition $\mathbf{NA}(\alpha)$, this proves the first part of the result.

The same argument shows that $\mathbb{P}\{\bar{r}(X,X') \neq r^*(X,X')\} \leq \mathbb{P}\{|\eta(X)-\eta(X')| < 2\|\eta-\bar{\eta}\|_\infty\}$. Combining this bound to $\mathbf{NA}(\alpha)$ permits to finish the proof when $q=\infty$.

When $q<\infty$, decompose $\mathcal{E}(\bar{r}) = \mathbb{E}[|\eta(X)-\eta(X')|\mathbb{I}\{(X,X')\in\Gamma_{\bar{\eta}}\}]$ into a sum of two terms, depending on whether $|\eta(X)-\eta(X')| \leq t$ or not. As above, the first term is bounded by $\mathbb{E}[2|\eta(X)-\bar{\eta}(X)|\mathbb{I}\{|\eta(X)-\eta(X')|\leq t\}]$. Combining Hölder inequality with $\mathbf{NA}(\alpha)$, one gets that $2\mathbb{E}[|\eta(X)-\hat{\eta}_n(X)|\mathbb{I}\{|\eta(X)-\eta(X')|<t\}]$ is bounded by $C^{(q-1)/q}t^{\alpha(q-1)/q}\|\eta-\bar{\eta}\|_q$. The second term is bounded by the expectation of

$$(|\eta(X)-\bar{\eta}(X)|+|\eta(X')-\bar{\eta}(X')|)\times \mathbb{I}\{|\eta(X)-\bar{\eta}(X)|+|\eta(X')-\bar{\eta}(X')|>t\},$$

which term can be shown to be smaller than $4\mathbb{E}[|\eta(X)-\bar{\eta}(X)|\mathbb{I}\{|\eta(X)-\bar{\eta}(X)|> t/2\}]$. Combining Hölder and Markov inequalities, this is bounded by $2^{q+1}\|\eta-\bar{\eta}\|_q^q/t^{q-1}$. Finally, minimizing in t, we obtain the desired result. The same argument can be applied to $\mathbb{P}\{\bar{r}(X,X')\neq r^*(X,X')\}$, in order to decompose it into two terms, whether $|\eta(X)-\eta(X')|\leq t$ or not. The first one is bounded by Ct^α and the other one by $2^{q+1}\|\eta-\bar{\eta}\|_q^q/t^q$. Hence, optimizing in t leads to the last bound stated in the Proposition.

Proof of Proposition 7.3.1

Proof. Using the convexity of the hinge loss, we have $A_n(\tilde{f}_n) \leq \sum_M^{j=1}\omega_j A_n(f_j)$. Let $j_0 = \arg\min_{j=1,\ldots,M} A_n(f_j)$, we have $A_n(f_j) = A_n(f_{j_0}) + \frac{1}{n}(\log(\omega_{j_0})-\log(\omega_j))$ for all $j=1,\ldots,M$ and by averaging over the ω_j, we obtain:

$$A_n(\tilde{f}_n) \leq \min_{j=1,\ldots,M} A_n(f_j) + \frac{1}{n}\sum_{j=1}^{M}\omega_j(\log(\omega_{j_0})-\log(\omega_j)),$$

Using that $\sum_M^{j=1}\omega_j\frac{\log\omega_j}{1/M} = K(w|u)\geq 0$ where $K(w|u)$ denotes the Kullback-Leiber divergence between the weights $\omega=(\omega_j)_{j=1,\ldots,M}$ and the uniform weights $u=(1/M)_{j=1,\ldots,M}$ and $w_{j_0}\leq 1$, we obtain the desired result. □

Proof of Theorem 7.3.2

Proof. Let $a>0$. Adding and subtracting $(1+a)(A_n(\tilde{f}_n)-A_n(f^*))$ to $A(\tilde{f}_n)-A^*$ and then using proposition 7.3.1, we have for any $f\in\mathcal{F}$:

$$A(\tilde{f}_n)-A^* \leq (1+a)(A_n(f)-A_n(f^*)+\frac{\log M}{n})+A(\tilde{f}_n)-A^*-(1+a)(A_n(\tilde{f}_n)-A_n(f^*)).$$

Taking the expectation, we upper bound $\mathbb{E}[A(\tilde{f}_n)-A^*]$ by

$$(1+a)\min_{f\in\mathcal{F}}(A(f)-A(f^*))+\frac{\log M}{n}+\mathbb{E}[A(\tilde{f}_n)-A^*-(1+a)(A_n(\tilde{f}_n)-A_n(f^*))].$$

Now the goal is to control the expectation in the RHS. For that we use the Bernstein's inequality. First, notice that, using the linearity of the hinge loss on $[-1,1]$ we have:

$$A(\tilde{f}_n) - A^* - (1+a)(A_n(\tilde{f}_n) - A_n(f^*)) \leq \max_{f \in \mathcal{F}} A(f) - A^* - (1+a)(A_n(f) - A_n(f^*)),$$

using the union bound we deduce that, for all $\delta \in]0, 4+2a[$, the probability $\mathbb{P}\{A(\tilde{f}_n) - A^* - (1+a)(A_n(\tilde{f}_n) - A_n(f^*))\} \geq \delta\}$ is bounded by the sum $\sum_{f \in \mathcal{F}} \mathbb{P}\{A(f) - A^* - (1+a)(A_n(f) - A_n(f^*)) \geq \delta\}$. The MA($\alpha$) assumption implies that the variance of $\mathbb{I}\{Z \neq f(X,X')\} - \mathbb{I}\{Z \neq f^*(X,X')\}$ is bounded by $(A(f) - A^*)^{\frac{\alpha}{1+\alpha}}$. Now, using the Bernstein's inequality on $\mathbb{P}\{A(f) - A^* - (A_n(f) - A_n(f^*)) \geq \frac{\delta + a(A(f)-A^*)}{1+a}\delta\}$, $\mathbb{P}\{A(\tilde{f}_n) - A^* - (1+a)(A_n(\tilde{f}_n) - A_n(f^*)) \geq \delta\}$ is bounded for all by $\delta \in]0, 4+2a[$

$$\sum_{f \in \mathcal{F}} \exp\left(- \frac{n(\delta + a(A(f) - A^*))^2}{2(1+a)^2 (A(f) - A^*)^{\frac{\alpha}{1+\alpha}} + 2(1+a)(\delta + a(A(f) - A^*))/3} \right).$$

The quantity insides the exponential is lower for all $\delta \in]0, 4+2a[$ and $f \in \mathcal{F}$ than $-c\delta^{2-\frac{\alpha}{1+\alpha}}$ where c depend only on a. Using the fact that $\int_u^{+\infty} \exp(-bt^\kappa)dt \leq \frac{\exp(-bu^\kappa)}{\kappa b u^{\kappa-1}}$ and the inequality obtained, we get

$$\mathbb{E}[A(\tilde{f}_n) - A^* - (1+a)(A_n(\tilde{f}_n) - A_n(f^*))] \leq 2t + M \frac{\exp(-nct^{\frac{2+\alpha}{2+2\alpha}})}{ncbt^{\frac{1}{1+\alpha}}}$$

Optimizing in t the RHS, we obtain the desired result. □

Proof of Theorem 7.4.1

Proof. Using Corollary 7.3.3 with $a = 1$, we get, for any $\varepsilon > 0$:

$$\mathcal{E}(r_{\tilde{f}^{\varepsilon,n}}) \leq 4 \min_{g \in \Lambda_\varepsilon(\beta)} (L(r_g - L^*) + C \left(\frac{\log \Lambda_\varepsilon(\beta)}{n} \right)^{\frac{\alpha+1}{\alpha+2}}.$$

Using that $L(r_g) - L^* \leq C\|g - \eta\|_{L^\infty}^{1+\alpha}$ (see [Clémençon & Robbiano, 2011]) we obtain that

$$\mathcal{E}(r_{\tilde{f}^{\varepsilon,n}}) \leq D \left(\varepsilon^{1+\alpha} + \left(\frac{\varepsilon^{-d/\beta}}{n} \right)^{\frac{\alpha+1}{\alpha+2}} \right).$$

Taking $\varepsilon_n = n^{-\alpha\beta/(d+\beta(2+\alpha))}$, we obtain the result. □

Proof of Theorem 7.4.2

Proof. We introduce the function $\phi :]0;1[\times]0;\infty[\to]0;1/2[, (\alpha,\beta) \mapsto \frac{\beta}{d+\beta(2+\alpha)}$. There exists n_1 depending on K such that for any n greater than n_1 we have, for all $(\alpha,\beta) \in K$

$$\ln(n)^{-1} \leq \phi(\alpha,\beta) \leq \lfloor \ln(n)/2 \rfloor \ln(n)^{-1}.$$

Let $(\alpha_0,\beta_0) \in K$. For $n \geq n_1$, we denote $a_0 \in \{1,\ldots,\lfloor \ln(n)/2 \rfloor\}$ the integer such that $\phi_{a_0} = a_0 \ln(n)^{-1} \leq \phi(\alpha_0,\beta_0) \leq (a_0+1)\ln(n)^{-1}$ and $q_0 \in \{1,\ldots,\lfloor \ln(n) \rfloor^2 - 1\}$ such that $\beta_{q_0} = q_0 \ln(n)^{-1} \leq \beta_0 \leq (q_0+1)\ln(n)^{-1}$. Denote by $g_{\beta_{q_0}}(.)$ the decreasing function $\phi(.,\beta_{q_0})$ from $[0,1]$ to $[0,1/2]$ and we set $\alpha_{0,n} = g^{-1}_{\beta_{q_0}}(\phi_{a_0})$. There exists A such that $A|\alpha_{0,n} - \alpha_0| \leq |g_{\beta_{q_0}}(\alpha_{0,n}) - g_{\beta_{q_0}}(\alpha_0)| \leq \ln(n)^{-1}$. Let P be a probability distribution belonging to $\mathcal{P}_{\alpha_0,\beta_0,\mu_{max}}$. Applying the Corollary 7.3.3 with $a = 1$, we get

$$\mathbb{E}\left[\mathcal{E}(r_{\tilde{f}^{adp}})|D^1_m\right] \leq 4 \min_{(\varepsilon,\beta)\in\mathcal{G}}(L(r) - L^*) + C\left(\frac{\ln \mathrm{Card}(\mathcal{G})}{l}\right)^{\frac{\alpha+1}{\alpha+2}}$$

Using that $l = n/\ln(n)$ and that $\mathrm{Card}(\mathcal{G}) \leq \ln(n)^3$ combined with the definition of a_0 and $\varepsilon^0_m = m^{-\frac{a_0}{\ln n}}$, we have

$$\mathbb{E}_P\left[\mathcal{E}(\tilde{r}^{adp})\right] \leq C\left(\mathbb{E}_P\left[\mathcal{E}\left(r^{\varepsilon^0_m,\beta_{a_0}}_m\right)\right] + C\left(\frac{\ln^2 n}{n}\right)^{\frac{\alpha+1}{\alpha+2}}\right)$$

where C is independent of n. Since $\beta_{a_0} \leq \beta_0$ and that there exists a constant A_1 such that $\alpha_0 > \alpha_{0,n} - A_1 \ln(n)^{-1} = \alpha'_{0,n}$, we have $\mathcal{P}_{\alpha_0,\beta_0,\mu_{max}} \subset \mathcal{P}_{\alpha'_{0,n},\beta_{a_0},\mu_{max}}$. Using theorem 7.4.1 we can upper bound $\mathbb{E}_P[\mathcal{E}(r^{\varepsilon^0_m\beta_{a_0}})]$ by $Cm^{-\psi(\alpha_0,\beta_{k_0})}$ where C depend on K and d and $\psi(\alpha,\beta) = \frac{\beta(1+\alpha)}{\beta(2+\alpha)+d}$. By construction, there exist A_2 such that $|\psi(\alpha_0,\beta_{a_0}) - \psi(\alpha_0,\beta_0)| \leq A_2 \ln(n)^{-1}$ and using that $n^{A_2/\ln(n)} = e^{A_2}$, we get

$$\mathbb{E}_P[\mathcal{E}(\tilde{r}^{adp})] \leq C\left(n^{-\psi(\alpha_0,\beta_0)} + C\left(\frac{\ln^2 n}{n}\right)^{\frac{\alpha_0+1}{\alpha_0+2}}\right)$$

We conclude the proof using that $\psi(\alpha_0,\beta_0) \leq \frac{\alpha_0+1}{\alpha_0+2}$. □

Proof of Theorem 7.4.3.

We start with establishing the following result.

Lemma 7.7.1. *Assume that condition* $\mathbf{NA}(\alpha)$ *holds for* $\alpha > 0$. *Let* $\widehat{\eta}_n$ *be an estimator of* η. *Assume that* $\mathcal{P}$ *is a set of joint distributions such that:* $\forall n \geq 1$,

$$\sup_{P\in\mathcal{P}} \mathbb{P}\left\{|\widehat{\eta}_n(X) - \eta(X)| > \delta\right\} \leq C_1 \exp(-C_2 a_n \delta^2), \tag{7.21}$$

for some constants C_1 *and* C_2. *Then, there exists a constant* $C < \infty$ *such that we have for all* $n \geq 1$*:*

$$\sup_{P\in\mathcal{P}} \mathbb{E}[\mathcal{E}(r_{\widehat{\eta}_n})] \leq C \cdot a_n^{(1+\alpha)/2}.$$

Proof. Let $u \in (0,1)$, consider the sequence of (disjoint) subsets of $\mathbb{R}^d$ defined by

$$A_0(u) = \{x \in \mathbb{R}^d : |\eta(x) - u| < \delta\},$$
$$A_j(u) = \{x \in \mathbb{R}^d : 2^{j-1}\delta < |\eta(x) - u| < 2^j\delta\}, \text{ for } j \geq 1.$$

For any $\delta > 0$, we may write $\mathcal{E}(r_{\widehat{\eta}_n})$ as

$$\sum_{j\geq 0} \mathbb{E}_{X'}\mathbb{E}_X[|\eta(X) - \eta(X')|\mathbb{I}\{X \in A_j(X')\}\mathbb{I}\{(X, X') \in \Gamma_{\widehat{\eta}_n}\}]$$

The term corresponding to $j = 0$ in the sum above is bounded by $C\delta^{1+\alpha}$ by virtue of assumption $\mathbf{NA}(\alpha)$. The one indexed by $j \geq 1$ is smaller than $2^{j+1}\delta\mathbb{E}\left[\mathbb{I}\{|\hat{\eta}_n(X) - \eta(X)| > 2^{j-2}\delta,\ X \in A_j(X')\}\right]$. Then, using the hypothesis on the class $\mathcal{P}$ plus assumption $\mathbf{NA}(\alpha)$, it is less than $2C_1 2^{j(1+\alpha)}\delta^{1+\alpha}\exp(-C_2 a_n(2^{j-2}\delta)^2)$. The proof is finished by summing all the bounds. □

It follows from Theorem 3.2 in [Audibert & A.Tsybakov, 2007] that (7.21) holds for the estimator considered with $a_n = n^{\frac{2\beta}{2\beta+d}}$ when $\mathcal{P} = \mathcal{P}_{\alpha,\beta,L}$. Using the lemma, this lead to the following upper bound for the excess risk. Now, using inequality (7.7) and taking $\delta = a_n^{-1/2}$, one gets the desired result.□

Proof of Theorem 7.4.4

Proof. We introduce the function $\Theta :]0;1[\times]0;\infty[\to]0;1/2[, (\alpha,\beta) \mapsto \frac{\beta(1+\alpha)}{d+2\beta}$. There exists n_1 depending on K such that for any n greater than n_1 we have, for all $(\alpha,\beta) \in K$

$$\min_{(\alpha,\beta)\in K}(1+\alpha)\ln(n)^{-1} \leq \Theta(\alpha,\beta) \leq \max_{(\alpha,\beta)\in K}(1+\alpha)\lfloor \ln(n)/2\rfloor \ln(n)^{-1}.$$

Let $(\alpha_0,\beta_0) \in K$ be such that $\alpha_0\beta_0 \leq d$. For $n \geq n_1$, we denote $k_0 \in \{1,\ldots,\lfloor \ln(n)/2\rfloor\}$ the integer such that

$$(1+\alpha_0)k_0\ln(n)^{-1} \leq \Theta(\alpha,\beta) \leq (1+\alpha_0)(k_0+1)\ln(n)^{-1}.$$

Let P be a probability distribution belonging to $\mathcal{P}_{\alpha_0,\beta_0,\mu_{max},\mu_{min}}$. Applying the Corollary 7.3.3 with $a = 1$, we get

$$\mathbb{E}\left[\mathcal{E}(r_{\tilde{f}^{adp}})|D_m^1\right] \leq 4\min_{r\in\mathcal{F}}(L(r) - L^*) + C\left(\frac{\ln \mathrm{Card}(\mathcal{F})}{l}\right)^{\frac{\alpha+1}{\alpha+2}}$$

Using that $l = n/\ln(n)$ and that $\mathrm{Card}(\mathcal{F}) \leq \ln(n)$ combined with the definition of k_0, we have

$$\mathbb{E}_P\left[\mathcal{E}(r_{\tilde{f}^{adp}})\right] \leq C\left(\mathbb{E}_P\left[\mathcal{E}\left(r_m^{\beta_{k_0}}\right)\right] + C\left(\frac{\ln^2 n}{n}\right)^{\frac{\alpha+1}{\alpha+2}}\right)$$

where C is independent of n. Since $\beta_{k_0} \leq \beta_0$, we have $\mathcal{P}_{\alpha_0,\beta_0,\mu_{max},\mu_{min}} \subset \mathcal{P}_{\alpha_0,\beta_{k_0},\mu_{max},\mu_{min}}$. Using theorem 7.4.3 we can upper bound $\mathbb{E}_P[\mathcal{E}(r_m^{\beta_{k_0}})]$ by $Cm^{-\Theta(\alpha_0,\beta_{k_0})}$ where C depend on K and d. By construction, we have $|\Theta(\alpha_0,\beta_{k_0}) - \Theta(\alpha_0,\beta_0)| \leq \ln(n)^{-1}$ and using that $n^{1/\ln(n)} = e$, we get

$$\mathbb{E}_P[\mathcal{E}(r_{\tilde{f}^{adp}})] \leq C\left(n^{-\Theta(\alpha_0,\beta_0)} + C\left(\frac{\ln^2 n}{n}\right)^{\frac{\alpha_0+1}{\alpha_0+2}}\right)$$

We conclude the proof using that $\Theta(\alpha_0,\beta_0) \leq \frac{\alpha_0+1}{\alpha_0+2}$ when $\alpha_0\beta_0 \leq d$. □

Proof of Theorem 7.5.1

Proof. The proof is classically based on Assouad's lemma. For $q \geq 1$, consider the regular grid on $[0;1]^d$ defined as

$$G^{(q)} = \left\{\left(\frac{2k_1+1}{2q}, \ldots, \frac{2k_d+1}{2q}\right) \text{ such as } k_1, \ldots, k_d \in \{0, ..., q-1\}\right\}.$$

Let $\eta_q(x) \in G^{(q)}$ be the closest point to $x \in [0;1]^d$ in $G^{(q)}$ (uniqueness of $\eta_q(x)$ is assumed: if it does not hold, define $\eta_q(x)$ as the one which is moreover closest to 0). Consider the partition $\mathcal{X}'_1, ..., \mathcal{X}'_{q^d}$ of $[0,1]^d$ canonically defined using the grid $G^{(q)}$(x and y belong to the same subset iff $\eta_q(x) = \eta_q(y)$). Obviously, $\mathcal{X} = [0,1]^d = \cup_{i=1}^{q^d} \mathcal{X}'_i$. Let $u_1 : \mathbb{R}_+ \to \mathbb{R}_+$ be a non increasing infinitely differentiable function as in [Audibert & A.Tsybakov, 2007]. Let $u_2 : \mathbb{R}_+ \to \mathbb{R}_+$ be an infinitely differentiable bump function such as $u'_2 = 1$ on $[1/12, 1/6]$. Let $\phi_1, \phi_2 : \mathbb{R}^d \to \mathbb{R}_+$ be function defined as

$$\phi_i(x) = C_\phi u_i(\|x\|), \tag{7.22}$$

where the positive constant C_ϕ is taken small enough to ensure that $|\phi_i(x) - \phi_{i,x}(x')| \leq L\|x' - x\|^\beta$ for any $x, x' \in \mathbb{R}$. Thus $\phi_1, \phi_2 \in \Sigma(\beta, L, \mathbb{R})$. Now we define the hypercube $\mathcal{H}$. For this purpose, we merge together intervals : $G_k = [(k-1)K/q; kK/q] \times [0,1]^{d-1}, k \in \{1, ..., H\}$ where K is the number of intervals we bring together relatively to the first coordinate (and it will play a role in the proof), $m = Kq^{d-1}$ is the number of cubes in a group and $H = \lfloor q/K \rfloor$. Define the hypercube $\mathcal{H} = \{\mathbb{P}_{\vec{\sigma}}, \vec{\sigma} \in S_m^H\}$, where S_m is the symmetric group of order m , of probability distributions $\mathbb{P}_{\vec{\sigma}}$ of (X,Y) as follows.

We design the marginal distribution of X that does not depend on σ and has a density μ w.r.t Lebesgue measure on $\mathbb{R}^d$. For fixed $0 < W$, we take μ as $\mu(x) = W/\lambda_d(B(z,1/4q))$ if x belongs to a set $B(z,1/6q) \setminus B(z,1/12q)$ for some $z \in G^{(q)}$, and $\mu(x) = 0$ for all other x. We call $\mathcal{X}_i = \mathcal{X}_i \cap B(z,1/6q) \setminus B(z,1/12q)$ for $i = 1, \ldots, m$. Next, the distribution of Y given X for $\mathbb{P}_{\sigma,k} \in \mathcal{H}$ is determined by the regression function, if $x \in \mathcal{X}'_i$ with $i \in \{1, \ldots, m\}$,

$$\eta_{\vec{\sigma}}(x) = k(x)K/q + \sigma^{k(x)}(x)\tilde{h}\phi_1\left(q|x - \eta_q(x)|\right) + \tilde{h}\phi_2\left(q|x - \eta_q(x)|\right).$$

where $\tilde{h}$ is a function of q and $k(x) = \lfloor xK/q \rfloor$.

We now check the assumptions. Because of the design Hölder condition holds for $x, x' \in \mathcal{X}_i$ ([Audibert & A.Tsybakov, 2007]). In contrast of classification situation, we have to check whether Hölder condition holds for $x \in \mathcal{X}_i, x' \in \mathcal{X}_j$ when $i \neq j$ belong to a same group G_k. One can see that Hölder condition holds as soon as $m\tilde{h} \leq Lq^{-\beta}$ (i.e $K\tilde{h} \leq Lq^{1-d-\beta}$) . Consider now the margin assumption. For $t = O(\tilde{h})$ the margin condition implies $W \leq C\tilde{h}^{\alpha}$. A constraint on K is also induced by the margin assumption: restricted to a group, the range of η has a measure of order $q^{-\beta}$ (because of the Hölder assumption). Hence, the margin assumption is satisfied if $mW = O(q^{-\alpha\beta})$ because of the strong density assumption $W \geq C/q^d$. Coupling the two last inequalities leads to $\alpha\beta \leq 1$, guaranteeing $K \geq 2$. So we take $\tilde{h} = C'q^{-d-\beta+\alpha\beta}$ and we verify that the margin condition holds. Indeed, if $\alpha\beta \leq d$, there exists $C' > 0$ such as $\tilde{h}^{\alpha} = C'q^{-\alpha d-\alpha\beta+\alpha^2\beta} \geq C/q^d$.

We denote $G(j)$, the set of cubes in the same group of j and for $i \in G(j), i \neq j$, $\sigma_{i,j} = +1$ if for all $x \in \mathcal{X}_i, x' \in \mathcal{X}_j$, $\eta_{\vec{\sigma}}(x) > \eta_{\vec{\sigma}}(x')$ and $\sigma_{i,j} = -1$ otherwise.

Using lemma 1, we have

$$\mathbb{E}_{\vec{\sigma}} L(r_n) - L^* = \frac{1}{2}\mathbb{E}_{\vec{\sigma}}\left[\mathbb{E}_{\vec{\sigma}}\left[\sum_{j=1}^{q^d} |\eta_{\vec{\sigma}}(X) - \eta_{\vec{\sigma}}(X')||r_{\vec{\sigma}}(X, X') - \hat{r}(X, X')|\mathbb{I}_{X' \in \mathcal{X}_j}\right]\right].$$

Using that $\mathcal{X} = \bigsqcup \mathcal{X}_i$ (i.e the disjoint union of the $\mathcal{X}_i$) combined with the definition of the margin law of X, we lower bound the excess risk by

$$\frac{W}{2}\mathbb{E}_{\vec{\sigma}}\left[\sum_{j=1}^{q^d} \mathbb{I}_{X' \in \mathcal{X}_j}\mathbb{E}_{\vec{\sigma}}\left[\sum_{i \in G(j), i \neq j} \int_{\mathcal{X}_i} |\eta_{\vec{\sigma}}(x) - \eta_{\vec{\sigma}}(X')||r_{\vec{\sigma}}(x, X') - \hat{r}(x, X')|\frac{dx}{\lambda(\mathcal{X}_i)}\right]\right].$$

We denote $d_\eta(\mathcal{X}_i, \mathcal{X}_j) = \min_{(x,x') \in (\mathcal{X}_i, \mathcal{X}_{j_0})} |\eta_\sigma(x) - \eta_\sigma(x')|$. Now, using the definition of $\sigma_{i,j}$ and $G(i)$, the last expression is lower than

$$\frac{W}{2}\mathbb{E}_{\vec{\sigma}}\left[\sum_{j=1}^{q^d} \mathbb{I}_{X' \in \mathcal{X}_j}\left[\mathbb{E}_{\vec{\sigma}}[\sum_{i \in G(j), i \neq j} d_\eta(\mathcal{X}_i, \mathcal{X}_j)\left|\sigma_{i,j} - \int_{\mathcal{X}_i} \hat{r}(x, X')\frac{dx}{\lambda(\mathcal{X}_i)}\right|\right]\right]$$

We denote by $\hat{\sigma}_{i,j} = \int_{X_i} r_n(x, X')\mathbb{I}_{X' \in \mathcal{X}_j}\frac{dx}{\lambda(\mathcal{X}_i)}$. So it remains to lower bound,

$$\sup_{\vec{\sigma} \in S_m^H} \mathbb{E}_{\vec{\sigma}}\left[\sum_{i \in G(j), i \neq j} d_\eta(\mathcal{X}_i, \mathcal{X}_j)|\sigma_{i,j} - \hat{\sigma}_{i,j}|\right].$$

Using that the sup is always greater than the mean and the linearity of the expectation, we lower bound by

$$\frac{1}{m!^H}\sum_{i \in G(j), i \neq j}\sum_{\vec{\sigma} \in S_m^H} \mathbb{E}_{\vec{\sigma}}\left[d_\eta(\mathcal{X}_i, \mathcal{X}_j)|\sigma_{i,j} - \hat{\sigma}_{i,j}|\right]$$

Restricting the sum to $\vec{\sigma}$'s such that the σ corresponding at the group $G(j)$ satisfies $\sigma(i) - \sigma(j) > m/2$ or $\sigma(j) - \sigma(i) > m/2$, we have $d_\eta(\mathcal{X}_i, \mathcal{X}_j) \geq C/q^\beta$. Combining this with the triangular inequality we obtain the following lower bound

$$\frac{1}{m!^H} \sum_{i \in G(j), i \neq j} \sum_{\vec{\sigma} \in S_m^H | \sigma_{i,j} = 1, \sigma(i) - \sigma(j) > m/2} \frac{1}{2q^\beta} \mathbb{E}_{\pi_\sigma, \pi_{\tau_{i,j}\sigma}} [|\sigma_{i,j} - \tau_{i,j}\sigma_{i,j}|]$$

where $\tau_{i,j}$ is the transposition (i, j). Using inequality between the divergence and the Hellinger's distance, the last expression is greater than

$$\frac{1}{m!^H} \sum_{i \in G(j), i \neq j} \sum_{\sigma \in S_m^H | \sigma_{i,j} = 1, \sigma(i) - \sigma(j) > m/2} \frac{1}{2q^\beta} (1 - \sqrt{1 - (1 - H^2(P_{\vec{\sigma}}^{\otimes n}, P_{\tau_{i,j}\vec{\sigma}}^{\otimes n})/2)^2}$$

A straightforward calculus shows that $H^2(P_{\vec{\sigma}}, P_{\tau_{i,j}\vec{\sigma}}) \leq 4W(1 - \sqrt{1 - q^{-2\beta}} \leq 4Wq^{-2\beta}$. Using argument of [Audibert, 2009] we have,

$$1 - \sqrt{1 - (1 - H^2(\mathbb{P}_\sigma^{\otimes n}, \mathbb{P}_{\tau_{i,i\_}\sigma}^{\otimes n})/2)^2}) \geq 1 - \sqrt{2n(W/q^{2\beta})}$$

The number of $\sigma \in S_m^H$ such that $\sigma(i) - \sigma(j) > m/2$ is greater than $m!^H/8$, so finally we have

$$\inf_{r_n} \sup_{P \in \Pr_{\alpha,\beta,\mu_{min},\mu_{max}}} L(r_n) - L^* \geq C \frac{Wm}{q^\beta} (1 - q^{-\beta}\sqrt{2nW})$$

Now, we take $q = C_1 n^{1/(2\beta+d)}$ combined with $W = C_2 q^{-d}$ and $m = C_3 q^{d-\alpha\beta}$ with some positive constants C_1, C_2, C_3, to conclude the proof. □

7.8 Annex - Discussion on the lower bounds

Basically, the idea of the proof of 7.5.1 is the following : first fix $\mathcal{X}_1 \subset \mathcal{X}$ a part of the space such that $X \in \mathcal{X}_1$ then create a classification problem as in [Audibert & A.Tsybakov, 2007] around $\mathcal{X}_1$. Doing that gives the rates of convergence of classification multiplied by the measure of $\mathcal{X}_1$. So the next step is to create classification problems for a union of part of the space with a measure independent of n. For the mild case in classification (see the proof in [Audibert & A.Tsybakov, 2007]), the classification problem uses the all space $\mathcal{X}$ i.e. all the important parts of the space have an η close to $1/2$ and a density close to zero. So with our strategy we obtain the rates of classification times the measure of the space $\mathcal{X}_1$ (i.e. W in the previous proof) which is not independent of n. For information only, we give the lower bounds that are achievable with this strategy. Since we believe they are not optimal, we do not give the proofs.

Oracle inequality Adapting the proof of Theorem 3 in [Lecué, 2008], one can get the next proposition. Let Pr_α be the set of all probability distributions such that $\mathrm{NA}(\alpha)$ holds.

Proposition 7.8.1. *(lower bound) For any integers M and n such that $M \leq \exp(n)$, there exist M prediction rules $f_1, \dots, f_M$ such that for any decision function $\hat{f}_n$ and any $a > 0$, we have*

$$\sup_{P \in \mathrm{Pr}_\alpha} \left[\mathbb{E}[L(\hat{f}_n) - L^*] - (1+a) \min_{j=1,\dots,M} (L(f_j) - L^*) \right] \geq C_1 \left(\frac{\log M}{n \log\log M} \right)^{\frac{2\alpha+2}{\alpha+2}},$$

where $C > 0$ is a constant depending only on α and c_0.

Notice that the power of n is half the power of n in the upper bound. Moreover a term in $\log\log(M)$ appears and comes from the fact that, we use permutations instead of the hypercube $\{-1,+1\}^{\log(M)}$.

Mild assumption In that case, using directly the same proof as in the strong case with the choice of the parameters as in [Audibert & A.Tsybakov, 2007], one can prove the following proposition.

Proposition 7.8.2. *Let $(\alpha, \beta) \in]0,1] \times \mathbb{R}_+^*$. There exists a constant $C > 0$ such that, for any ranking rule r_n based on n independent copies of the pair (X, Y), we have: $\forall n \geq 1$,*

$$\sup_{P \in \mathrm{Pr}_{\alpha,\beta,\mu_{max},\mu_{min}}} \mathcal{E}(r_n) \geq C \cdot n^{-\frac{\beta(1+2\alpha)}{d+(2+\alpha)\beta}}.$$

Notice that the only change here is the factor 2 in front of the α.

Bibliography

[Agarwal, 2008] Agarwal, S. (2008). Generalization bounds for some ordinal regression algorithms. in *Proceedings of ALT'08*. (Cité en pages 40 et 41.)

[Agarwal *et al.*, 2005] Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., & Roth, D. (2005). Generalization bounds for the Area Under the ROC Curve. Journal of Machine Learning Research *6*, 393–425. (Cité en pages 2, 21, 28, 78 et 149.)

[Allwein *et al.*, 2001] Allwein, E., Schapire, R., & Singer, Y. (2001). Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of Machine Learning Research *1*, 113–141. (Cité en pages 15 et 92.)

[Alquier & Lounici, 2011] Alquier, P. & Lounici, K. (2011). Pac-bayesian bounds for sparse regression estimation with exponential weights. Electronic Journal of Statistics *5*, 127145. (Cité en page 156.)

[Audibert, 2009] Audibert, J. (2009). Fast learning rates in statistical inference through aggregation. The Annals of Statistics *37*, 1591–1646. (Cité en pages 147 et 171.)

[Audibert & A.Tsybakov, 2007] Audibert, J.-Y. & A.Tsybakov (2007). Fast learning rates for plug-in classifiers. The Annals of Statistics *35*(2), 608–633. (Cité en pages 22, 24, 52, 147, 148, 152, 155, 158, 161, 168, 169, 170 et 172.)

[Barthélemy *et al.*, 1989] Barthélemy, J., Guénoche, A., & Hudry, O. (1989). Median linear orders: heuristics and a branch and bound algorithm. European Journal of Operational Research *42*(3), 313–325. (Cité en page 97.)

[Bartlett *et al.*, 2006] Bartlett, P. L., Jordan, M. I., & McAuliffe, J. D. (2006). Convexity, classification and risk bounds. Journal of the American Statistical Association *101*, 138–156. (Cité en pages 22 et 151.)

[Baskiotis *et al.*, 2010] Baskiotis, N., Clémençon, S., Depecker, M., & Vayatis, N. (2010). R-implementation of the TreeRank algorithm. in *Proceedings of SMDTA'2010*. (Cité en pages 109 et 133.)

[Bertail *et al.*, 2008] Bertail, P., Clémençon, S., & Vayatis., N. (2008). On bootstrapping the roc curve. in *Proceedings of NIPS'08* pp. 137–144. (Cité en pages 10, 62 et 68.)

[Bertail & Tressou, 2006] Bertail, P. & Tressou, J. (2006). Incomplete generalized U-statistics for food risk assessment. Biometrics *62*(1), 66–74. (Cité en page 81.)

[Beygelzimer *et al.*, 2005a] Beygelzimer, A., Dani, V., Hayes, T., Langford, J., & Zadrozny, B. (2005a). Reductions between classification tasks. in *Proceedings of ICML'05*. (Cité en pages 15 et 92.)

[Beygelzimer *et al.*, 2005b] Beygelzimer, A., Langford, J., & Zadrozny, B. (2005b). Weighted one against all. in *Proceedings of AAAI'05*. (Cité en pages 15 et 92.)

[Biau & Bleakley, 2006] Biau, G. & Bleakley, L. (2006). Statistical inference on graphs. Statistics & Decisions *24*, 209–232. (Cité en page 77.)

[Blom, 1976] Blom, G. (1976). Some properties of incomplete U-statistics. Biometrika *63*(3), 573–580. (Cité en pages 12, 77 et 80.)

[Boucheron *et al.*, 2005] Boucheron, S., Bousquet, O., & Lugosi, G. (2005). Theory of classification: a survey of some secent advances. ESAIM: Probability and Statistics *9*, 323–375. (Cité en pages 77, 88 et 155.)

[Breiman *et al.*, 1984] Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees* (Wadsworth and Brooks: London). (Cité en page 109.)

[Brown & Kildea, 1978] Brown, B. & Kildea, D. (1978). Reduced U-statistics and the Hodges-Lehmann estimator. The Annals of Statistics *6*, 828–835. (Cité en page 81.)

[Cesa-Bianchi & Lugosi, 2006] Cesa-Bianchi, N. & Lugosi, G. (2006). *Prediction, Learning, and Games* (Cambridge University Press: New York). (Cité en page 156.)

[Chapelle & Chang, 2011] Chapelle, O. & Chang, Y. (2011). Yahoo! learning to rank challenge overview. Journal of Machine Learning Research *14*, 1–24. (Cité en pages 3 et 29.)

[Charon & Hudry, 1998] Charon, I. & Hudry, O. (1998). Lamarckian genetic algorithms applied to the aggregation of preferences. Annals of Operations Research *80*, 281–297. (Cité en page 97.)

[Clémençon, 2011] Clémençon, S. (2011). On U-processes and clustering performance. in *NIPS'11* pp. 37–45. (Cité en page 77.)

[Clémençon *et al.*, 2011a] Clémençon, S., Depecker, M., & Vayatis, N. (2011a). Adaptive partitioning schemes for bipartite ranking. Machine Learning *43*(1), 31–69. (Cité en pages 18, 47, 69, 94, 104, 109, 110, 111, 116 et 133.)

[Clémençon *et al.*, 2011b] Clémençon, S., Depecker, M., & Vayatis, N. (2011b). Avancées récentes dans le domaine de l'apprentissage statistique d'ordonnancements. Revue d'Intelligence Artificielle *25*(3), 345–368. (Cité en page 18.)

[Clémençon *et al.*, 2012] Clémençon, S., Depecker, M., & Vayatis, N. (2012). An empirical comparison of learning algorithms for nonparametric scoring: the TREERANK algorithm and other methods. Pattern Analysis and Applications. (Cité en page 111.)

[Clémençon *et al.*, 2013a] Clémençon, S., Depecker, M., & Vayatis, N. (2013a). Ranking forests. Journal of Machine Learning Research *14*(1), 39–73. (Cité en pages 4 et 29.)

[Clémençon *et al.*, 2005] Clémençon, S., Lugosi, G., & Vayatis, N. (2005). Ranking and scoring using empirical risk minimization. in *Proceedings of COLT*. (Cité en page 21.)

[Clémençon *et al.*, 2008] Clémençon, S., Lugosi, G., & Vayatis, N. (2008). Ranking and empirical risk minimization of U-statistics. The Annals of Statistics *36*(2), 844–874. (Cité en pages 2, 9, 17, 22, 28, 51, 57, 77, 78, 80, 83, 86, 95, 99, 100, 105, 117, 147, 149, 150, 151, 153 et 154.)

[Clémençon & N.Vayatis, 2009] Clémençon, S. & N.Vayatis (2009). Adaptive estimation of the optimal roc curve and a bipartite ranking algorithm. in *Proceedings of ALT'09*. (Cité en page 94.)

[Clémençon & Robbiano, 2011] Clémençon, S. & Robbiano, S. (2011). Minimax learning rates for bipartite ranking and plug-in rules. in *Proceedings of ICML'11*. (Cité en pages 162 et 166.)

[Clémençon *et al.*, 2013b] Clémençon, S., Robbiano, S., & N.Vayatis (2013b). Ranking data with ordinal labels: Optimality and pairwise aggregation. Machine Learning *91*(1), 67–104. (Cité en pages 111, 117, 120 et 121.)

[Clémençon & Vayatis, 2007] Clémençon, S. & Vayatis, N. (2007). Ranking the best instances. Journal of Machine Learning Research *8*, 2671–2699. (Cité en pages 2, 20, 28 et 132.)

[Clémençon & Vayatis, 2008] Clémençon, S. & Vayatis, N. (2008). Empirical performance maximization based on linear rank statistics. in *Proceedings of NIPS'08*. Springer. (Cité en pages 2 et 28.)

[Clémençon & Vayatis, 2009a] Clémençon, S. & Vayatis, N. (2009a). On partitioning rules for bipartite ranking. in *Proceedings of AISTATS* volume 5 pp. 97–104. Journal of Machine Learning Research: W&CP. (Cité en pages 52, 94, 148 et 155.)

[Clémençon & Vayatis, 2009b] Clémençon, S. & Vayatis, N. (2009b). Tree-based ranking methods. IEEE Transactions on Information Theory *55*(9), 4316–4336. (Cité en pages 2, 6, 16, 17, 18, 28, 35, 37, 42, 51, 95, 98, 104, 109, 110, 111, 114, 117, 122 et 133.)

[Clémençon & Vayatis, 2010] Clémençon, S. & Vayatis, N. (2010). Overlaying classifiers: a practical approach to optimal scoring. Constructive Approximation *32*(3), 619–648. (Cité en pages 2, 3, 28, 29, 51, 94, 98, 111, 115 et 150.)

[Cortes & Mohri, 2004] Cortes, C. & Mohri, M. (2004). AUC optimization vs. error rate minimization. in *Proceedings of NIPS'04*. (Cité en page 61.)

[Cossock & Zhang, 2008] Cossock, D. & Zhang, T. (2008). Statistical analysis of bayes optimal subset ranking.. IEEE Transactions on Information Theory *54*, 51405154. (Cité en pages 3 et 29.)

[Csorgo & Revesz, 1981] Csorgo, M. & Revesz, P. (1981). *Strong approximations en probability and statistics* (Academic press: London). (Cité en pages 10, 65 et 73.)

[Dalalyan & Tsybakov, 2008] Dalalyan, A. & Tsybakov, A. B. (2008). Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. Machine Learning *72*(1-2), 39–61. (Cité en page 156.)

[David, 2008] David, A. B. (2008). Ordinal real-world data sets repository. (Cité en pages 69 et 131.)

[Debnath *et al.*, 2004] Debnath, R., Takahide, N., & Takahashi, H. (2004). A decision based one-against-one method for multi-class support vector machine. Pattern Analysis and its Applications *7*, 164–175. (Cité en pages 15 et 92.)

[Depecker, 2010] Depecker, M. (2010). Méthodes d'apprentissage statistique pour le scoring. Master's thesis Télécom Paristech. (Cité en page 48.)

[Devroye *et al.*, 1996] Devroye, L., Györfi, L., & Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition* (Springer: New York). (Cité en pages 117, 120 et 121.)

[Dietterich & Bakiri, 1995] Dietterich, T. G. & Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. Journal of Articial Intelligence Research *2*, 263–286. (Cité en pages 15 et 92.)

[Dreiseitl *et al.*, 2000] Dreiseitl, S., Ohno-Machado, L., & Binder, M. (2000). Comparing three-class diagnostic tests by three-way ROC analysis. Medical Decision Making *20*, 323–331. (Cité en page 27.)

[Dudley, 1999] Dudley, R. (1999). *Uniform Central Limit Theorems* (Cambridge University Press: London). (Cité en pages 81 et 153.)

[Dvoretzky *et al.*, 1956] Dvoretzky, A., Kiefer, J., & Wolfowitz (1956). Asymptotic minimmax character of the sample distribution and the classical multinomial estimator. The Annals Mathematical statistics *27*, 642–669. (Cité en page 73.)

[Edwards *et al.*, 2005] Edwards, D., Metz, C., & Kupinski, M. (2005). The hypervolume under the roc hypersurface of 'near-guessing' and 'near-perfect' observers in n-class classification tasks.. IEEE Transactions on Medical Imaging *24*, 293–299. (Cité en page 27.)

[Efron, 1979] Efron, B. (1979). Bootstrap methods: another look at the jacknife. Annals of Statistics *7*, 1–26. (Cité en pages 10 et 66.)

[Enqvist, 1978] Enqvist, E. (1978). *On sampling from sets of random variables with application to incomplete U-statistics* PhD thesis. (Cité en page 81.)

[Fagin *et al.*, 2003] Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., & Vee, E. (2003). Comparing and aggregating rankings with ties. in *Proceedings of PODS'04* pp. 366–375. (Cité en page 97.)

[Falk & Reiss, 1989] Falk, M. & Reiss, R. (1989). Weak convergence of smoothed and nonsmoothed bootstrap quantile estimates. Annals of Probability *17*, 362–371. (Cité en pages 61 et 66.)

[Fawcett, 2006] Fawcett, T. (2006). An Introduction to ROC Analysis. Letters in Pattern Recognition *27*(8), 861–874. (Cité en pages 7 et 61.)

[Ferri *et al.*, 2002] Ferri, C., Flach, P., & Hernández-Orallo, J. (2002). Learning decision trees using the area under the ROC curve. in *Proceedings of ICML'02* pp. 139–146 San Francisco, CA, USA. (Cité en page 109.)

[Ferri *et al.*, 2003] Ferri, C., Hernández-Orallo, J., & Salido, M. (2003). Volume under the ROC surface for multi-class problems. in *Proceedings of ECML'03*. (Cité en pages 8, 9, 43, 46 et 50.)

[Fieldsend & Everson, 2005] Fieldsend, J. & Everson, R. (2005). Formulation and comparison of multi-class ROC surfaces. in *Proceedings of ROCML 2005*. (Cité en pages 8 et 43.)

[Fieldsend & Everson, 2006] Fieldsend, J. & Everson, R. (2006). Multi-class ROC analysis from a multi-objective optimisation perspective. Pattern Recognition Letters *27*, 918–927. (Cité en pages 8 et 43.)

[Fisher, 1936] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics *7*, 179–188. (Cité en pages 2 et 28.)

[Flach, 2004] Flach, P. (2004). Tutorial: "the many faces of ROC analysis in machine learning". part iii Technical report ICML'04. (Cité en pages 2, 8, 28 et 43.)

[Frank & Asuncion, 2010] Frank, A. & Asuncion, A. (2010). UCI machine learning repository. (Cité en page 130.)

[Frank & Hall, 2001] Frank, E. & Hall, M. (2001). A simple approach to ordinal classification. in *Proceeding of ECML'01*. (Cité en pages 40, 91, 92, 103 et 133.)

[Freund *et al.*, 2003] Freund, Y., Iyer, R. D., Schapire, R. E., & Singer, Y. (2003). An efficient boosting algorithm for combining preferences. Journal of Machine Learning Research *4*, 933–969. (Cité en pages 2, 3, 21, 28, 111, 135 et 149.)

[Freund & Schapire, 1999] Freund, Y. & Schapire, R. E. (1999). Short introduction to boosting. Journal of Japanese Society for Artificial Intelligence *14*, 771–782. (Cité en pages 3 et 28.)

[Friedman *et al.*, 1998] Friedman, J., Hastie, T., & Tibshirani, R. (1998). Additive logistic regression: a statistical view of boosting. The Annals of Statistics *28*, 2000. (Cité en pages 2 et 28.)

[Fürnkranz, 2002] Fürnkranz, J. (2002). Round robin classification. Journal of Machine Learning Research *2*, 721–747. (Cité en page 91.)

[Fürnkranz *et al.*, 2009] Fürnkranz, J., Hüllermeier, E., & Vanderlooy, S. (2009). Binary decomposition methods for multipartite ranking. in *ECML PKDD '09*. (Cité en pages 91, 92, 103 et 132.)

[Giné & Guillou, 2002] Giné, E. & Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimator. Annales Institut Poincaré (B), Probabilités et Statistiques *38*, 907–921. (Cité en page 74.)

[Giné & Zinn, 1984] Giné, E. & Zinn, J. (1984). Some limit theorems for empirical processes. The Annals of Probability *12*(4), 929–989. (Cité en page 87.)

[Green & Swets, 1966] Green, D. & Swets, J. (1966). *Signal detection theory and psychophysics* (Wiley: New York). (Cité en pages 2, 6, 28 et 41.)

[Gu & Ghosal, 2008] Gu, J. & Ghosal, S. (2008). Strong approximations for resample quantile processes and application to ROC methodology. Journal of Nonparametric Statistics *20*, 229–240. (Cité en page 9.)

[Hall *et al.*, 2004] Hall, P., Hyndman, R., & Fan, Y. (2004). Nonparametric confidence intervals for receiver operating characteristic curves. Biometrika *91*, 743–750. (Cité en pages 9, 61 et 68.)

[Hand & Till, 2001] Hand, D. & Till, R. (2001). A simple generalization of the area under the ROC curve for multiple class classification problems. Machine Learning *45*(2), 171–186. (Cité en pages 8, 43, 48 et 132.)

[Hanley & McNeil, 1982] Hanley, J. & McNeil, J. (1982). The meaning and use of the area under a ROC curve. Radiology (143), 29–36. (Cité en page 147.)

[Hastie & Tibshirani, 1998] Hastie, T. & Tibshirani, R. (1998). Classification by pairwise coupling. The Annals of Statistics *26*(2), 451–471. (Cité en pages 15, 91 et 92.)

[Hastie *et al.*, 2001] Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning* (Springer-Verlag: New York). (Cité en pages 2 et 28.)

[Hastie & Tibshirani, 1990] Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized additive models* (Chapman & Hall: London). (Cité en pages 2 et 28.)

[Herbrich *et al.*, 2000] Herbrich, R., Graepel, T., & Obermayer, K. (2000). Large margin rank boundaries for ordinal regression in *Advances in Large Margin Classifiers* (MIT Press). (Cité en pages 31, 83, 132 et 135.)

[Higgins, 2004] Higgins, J. (2004). *Introduction to Modern Nonparametric Statistics* (Duxbury Press: New York). (Cité en page 132.)

[Hoeffding, 1948] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. Ann. Math. Stat. *19*, 293–325. (Cité en pages 12, 77, 80 et 87.)

[Hsieh & Turnbull, 1996] Hsieh, D. & Turnbull, B. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. The Annals of Statistics *24*, 25–40. (Cité en pages 9 et 61.)

[Hudry, 2008] Hudry, O. (2008). NP-hardness results for the aggregation of linear orders into median orders. Annals of Operations Research *163*, 63–88. (Cité en page 97.)

[Huhn & Hüllermeier, 2009] Huhn, J. & Hüllermeier, E. (2009). Is an ordinal class structure useful in classifier learning?. International Journal of Data Mining, Modelling and Management *1*, 45–67(23. (Cité en page 93.)

[Hüllermeier *et al.*, 2008] Hüllermeier, E., Fürnkranz, J., Cheng, W., & Brinker, K. (2008). Label ranking by learning pairwise preferences. Artificial Intelligence *172*, 1897–1917. (Cité en page 111.)

[Janson, 1984] Janson, S. (1984). The asymptotic distributions of incomplete U-statistics. Z. Wahrsch. verw. Gebiete *66*, 495–505. (Cité en pages 80 et 81.)

[Joachims, 2002] Joachims, T. (2002). Optimizing search engines using clickthrough data. in *Proceedings of KDD'02* pp. 133–142. (Cité en pages 3, 20 et 28.)

[Kolmogorov & Tikhomirov, 1961] Kolmogorov, A. N. & Tikhomirov, V. M. (1961). ε-entropy and ε-capacity of sets in functional spaces.. American Mathematical Society Translations: Series 2, *17*, 277–364. (Cité en pages 24 et 159.)

[Koltchinskii & Beznosova, 2005] Koltchinskii, V. & Beznosova, O. (2005). Exponential convergence rates in classification. in *Proceedings of COLT'05*. (Cité en page 147.)

[Kramer *et al.*, 2001] Kramer, S., Pfahringer, B., Widmer, G., & Groeve, M. D. (2001). Prediction of ordinal regression trees. Fundamenta Informaticae *47*, 1001–1013. (Cité en page 41.)

[Laguna *et al.*, 1999] Laguna, M., Marti, R., & Campos, V. (1999). Intensification and diversification with elite tabu search solutions for the linear ordering problem. Computers and Operations Research *26*(12), 1217–1230. (Cité en page 97.)

[Landgrebe & Duin, 2006] Landgrebe, T. & Duin, R. (November 2006). A simplified extension of the area under the ROC to the multiclass domain. in *Proceedings of PRASA'06*. (Cité en pages 9 et 50.)

[Lebanon & Lafferty, 2003] Lebanon, G. & Lafferty, J. (2003). Conditional models on the ranking poset. in *Proceedings of NIPS'03*. (Cité en page 97.)

[Lecué, 2006] Lecué, G. (2006). Optimal oracle inequality for aggregation of classifiers under low noise condition. in *Proceedings of COLT'06*. (Cité en pages 22, 23, 24, 147, 151, 156, 158 et 159.)

[Lecué, 2008] Lecué, G. (2008). Classification with minimax fast rates for classes of bayes rules with sparse representation. Electronic Journal of Statistics *2*, 741–773. (Cité en pages 147, 155 et 172.)

[Ledoux & Talagrand, 1991] Ledoux, M. & Talagrand, M. (1991). *Probability in Banach Spaces* (Springer: New York). (Cité en page 77.)

[Lee, 1990] Lee, A. J. (1990). *U-statistics: Theory and practice* (Marcel Dekker, Inc.: New York). (Cité en pages 77, 79 et 81.)

[Lehmann & Romano, 2005] Lehmann, E. & Romano, J. P. (2005). *Testing Statistical Hypotheses* (Springer: New York). (Cité en pages 3, 29, 38 et 45.)

[Lepski *et al.*, 1997] Lepski, O., E.Mammen, & Spokoiny, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. The Annals of Statistics *25*, 929–947. (Cité en page 147.)

[Li & Zhou, 2009] Li, J. & Zhou, X. (2009). Nonparametric and semiparametric estimation of the three way receiver operating characteristic surface. Journal of Statistical Planning and Inference *139*, 4133–4142. (Cité en pages 9, 10, 35, 57, 61, 64, 65 et 72.)

[Liu *et al.*, 2007] Liu, T., Xu, J., Qin, T., Xiong, W., & Li, H. (2007). LETOR: Benchmark dataset for research on learning to rank for information retrieval. in *Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval* pp. 3–10. (Cité en page 78.)

[Macskassy & Provost, 2004] Macskassy, S. & Provost, F. (2004). Confidence bands for ROC curves: Methods and an empirical study. in *First Workshop on ROC Analysis in AI*. (Cité en pages 9 et 61.)

[Mandhani & Meila, 2009] Mandhani, B. & Meila, M. (2009). Tractable search for learning exponential models of rankings. in *Proceedings of AISTATS, Vol. 5 of Journal of Machine Learning Research:W&CP 5*. (Cité en page 97.)

[Massart, 2000] Massart, P. (2000). Some applications of concentration inequalities to statistics. Les Annales de la Faculté des Sciences de Toulouse *9*, 245–303. (Cité en pages 23, 147 et 154.)

[Massart, 2006] Massart, P. (2006). *Concentration inequalities and model selection* (Springer). (Cité en page 157.)

[Massart & Nédélec, 2006] Massart, P. & Nédélec, E. (2006). Risk bounds for statistical learning. The Annals of Statistics *34*(5). (Cité en page 150.)

[McDiarmid, 1989] McDiarmid, C. (1989). On the method of bounded differences. in *Surveys in Combinatorics* (Cambridge University Press). (Cité en page 87.)

[Meila *et al.*, 2007] Meila, M., Phadnis, K., Patterson, A., & Bilmes, J. (2007). Consensus ranking under the exponential model. in *Proceedings of UAI'07* pp. 729–734. (Cité en page 97.)

[Monnier, 2012] Monnier, J.-B. (2012). Classification via local multi-resolution projections. Electronic Journal of Statistics *6*, 382–420. (Cité en page 147.)

[Mossman, 1999] Mossman, D. (1999). Three-way ROCs. Medical Decision Making *78*, 78–89. (Cité en pages 1 et 27.)

[Nakas & Yiannoutsos, 2004] Nakas, C. & Yiannoutsos, C. (2004). Ordered multiple-class ROC analysis with continuous measurements. Statistics in Medicine *23*, 3437–3449. (Cité en pages 1, 27 et 61.)

[Pahikkala *et al.*, 2007] Pahikkala, T., Tsivtsivadze, E., Airola, A., Boberg, J., & Salakoski, T. (2007). Learning to rank with pairwise regularized least-squares. in *Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval* pp. 27–33. (Cité en pages 3, 20, 28, 31, 111 et 136.)

[Pepe, 2003] Pepe, M. (2003). *Statistical Evaluation of Medical Tests for Classification and Prediction* (Oxford University Press: Oxford). (Cité en pages 1, 7, 27 et 61.)

[Provost & Domingos, 2003] Provost, F. & Domingos, P. (2003). Tree induction for probability-based ranking. Machine Learning *52*(3), 199–215. (Cité en page 109.)

[Rajaram & Agarwal, 2005] Rajaram, S. & Agarwal, S. (2005). Generalization bounds for k-partite ranking. in *NIPS'05 Workshop on Learning to Rank*. (Cité en pages 51 et 111.)

[Rigollet & Tsybakov, 2011] Rigollet, P. & Tsybakov, A. (2011). Sparse estimation by exponential weighting. Statistical Science *27*, 558–575. (Cité en page 156.)

[Robbiano, 2013] Robbiano, S. (2013). Consistent aggregation of bipartite scoring functions for ranking ordinal data Technical report HAL. (Cité en page 136.)

[Rudin, 2006] Rudin, C. (2006). Ranking with a P-Norm Push. in *Proceedings of COLT'06*. (Cité en pages 2 et 28.)

[Rudin *et al.*, 2005] Rudin, C., Cortes, C., Mohri, M., & Schapire, R. E. (2005). Margin-based ranking and boosting meet in the middle. in *Proceedings of COLT'05*. (Cité en pages 2, 28 et 31.)

[Scurfield, 1996] Scurfield, B. (1996). Multiple-event forced-choice tasks in the theory of signal detectability. Journal of Mathematical Psychology *40*, 253–269. (Cité en pages 2, 8, 28, 35, 48 et 49.)

[Shorack & Wellner, 1986] Shorack, G. & Wellner, J. (1986). *Empirical processes with applications to statistics* (Wiley: New York). (Cité en page 74.)

[Silverman & Young, 1987] Silverman, B. & Young, G. (1987). The bootstrap: to smooth or not to smooth?. Biometrika *74*, 469–479. (Cité en page 64.)

[Silverman & Green, 1986] Silverman, B. W. & Green, P. J. (1986). *Density Estimation for Statistics and Data Analysis* (Chapman and Hall: London). (Cité en page 68.)

[Srebro *et al.*, 2010] Srebro, N., Sridharan, K., & Tewari, A. (2010). Smoothness, low noise and fast rates. in *Proceedings of NIPS'10*. (Cité en page 147.)

[Stone, 1982] Stone, C. (1982). Optimal global rates of convergence for nonparametric regression. The Annals of Statistics *10*, 1040–1053. (Cité en page 156.)

[Tsybakov, 2004] Tsybakov, A. (2004). Optimal aggregation of classifiers in statistical learning. The Annals of Statistics *32*(1), 135–166. (Cité en pages 17, 23, 99, 147, 150 et 153.)

[Vapnik, 1999] Vapnik, V. (1999). An overview of statistical learning theory. IEEE Transactions on Neural Networks *10*(5), 988–999. (Cité en page 51.)

[Venkatesan & Amit, 1999] Venkatesan, G. & Amit, S. (1999). Multiclass learning, boosting, and error-correcting codes. in *Proceedings of COLT '99*. (Cité en pages 15 et 92.)

[Voorhees & Harman, 2005] Voorhees, E. M. & Harman, D. K. (2005). *TREC: Experiment and Evaluation in Information Retrieval* (MIT Press). (Cité en pages 3 et 29.)

[Waegeman & Baets, 2011] Waegeman, W. & Baets, B. D. (2011). On the era ranking representability of pairwise bipartite ranking functions. Artificial Intelligence *175*, 1223–1250. (Cité en pages 5, 6, 30, 35 et 38.)

[Waegeman *et al.*, 2008a] Waegeman, W., Baets, B. D., & Boullart, L. (2008a). Learning layered ranking functions with structured support vector machines. Neural Networks *21*, 1511–1523. (Cité en pages 3, 29 et 40.)

[Waegeman *et al.*, 2008b] Waegeman, W., Baets, B. D., & Boullart, L. (2008b). On the scalability of ordered multi-class ROC analysis. Computational Statistics and Data Analysis *52*, 3371–3388. (Cité en pages 50 et 51.)

[Waegeman *et al.*, 2008c] Waegeman, W., Baets, B. D., & Boullart, L. (2008c). ROC analysis in ordinal regression learning. Pattern Recognition Letters *29*, 1–9. (Cité en pages 8, 9, 43, 50 et 111.)

[Wakabayashi, 1998] Wakabayashi, Y. (1998). The complexity of computing medians of relations. Resenhas *3*(3), 323–349. (Cité en page 97.)

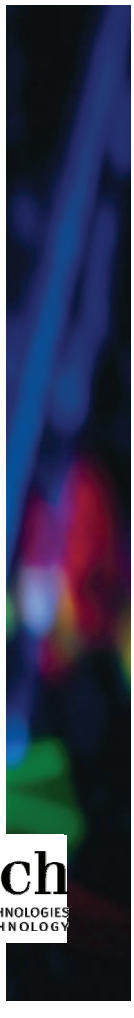
[Xia *et al.*, 2006] Xia, F., Zhang, W., & Wang, J. (2006). An effective tree-based algorithm for ordinal regression. IEEE Intelligent Informatics Bulletin *7*(1), 22–26. (Cité en page 109.)

[Xu & Li, 2007] Xu, J. & Li, H. (2007). AdaRank: a boosting algorithm for information retrieval. in *Proceedings of SIGIR'07* pp. 391–398. (Cité en pages 3, 20 et 28.)

[Yang, 1999] Yang, Y. (1999). Minimax nonparametric classification. I. rates of convergence. II. model selection for adaptation. IEEE Transactions on Information Theory *45*, 2271–2292. (Cité en page 156.)

[Zhang, 2004] Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization (with discussion). The Annals of Statistics *32*, 56–85. (Cité en pages 22 et 151.)

[Zhu & Hastie, 2001] Zhu, J. & Hastie, T. (2001). Kernel logistic regression and the import vector machine. in *Journal of Computational and Graphical Statistics* pp. 1081–1088. (Cité en pages 2 et 28.)



TELECOM ParisTech

EDiF

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY